Deep learning for image recognition II

Data Science Retreat Berlin | 28.02.2017

Ludwig Schmidt-Hackenberg

Quiz I

Q: What does the weights vector do?

A: It defines the importance of a neuron's inputs

Q: What does the bias do?

A: It changes the neuron's output independent of the input

Q: What is gradient descent?

A: A optimization algorithm to find a local minimum for a function

Quiz II

Q: What is the learning rate?

A: Is the step size (factor) of the gradient descent

Q: What is stochastic gradient descent?

A: Using only a random (thus stochastic) subset of the data for gradient descent

Q: What is an epoch?

A: A SGD training run using all data

Quiz III

Q: What is the cost function?

A: A function that returns a number representing how well the neural network performed mapping training examples to correct output.

Q: What is a loss function?

A: A different name for the cost function.

Q: What is regularization?

A: (Here) The process of introducing additional information into a learning process to prevent overfitting.

Quiz III

Q: What is the cost function?

A: A function that returns a number representing how well the neural network performed mapping training examples to correct output.

Q: What is a loss function?

A: A different name for the cost function.

Q: What is regularization?

A: (Here) The process o
a learning process to pre

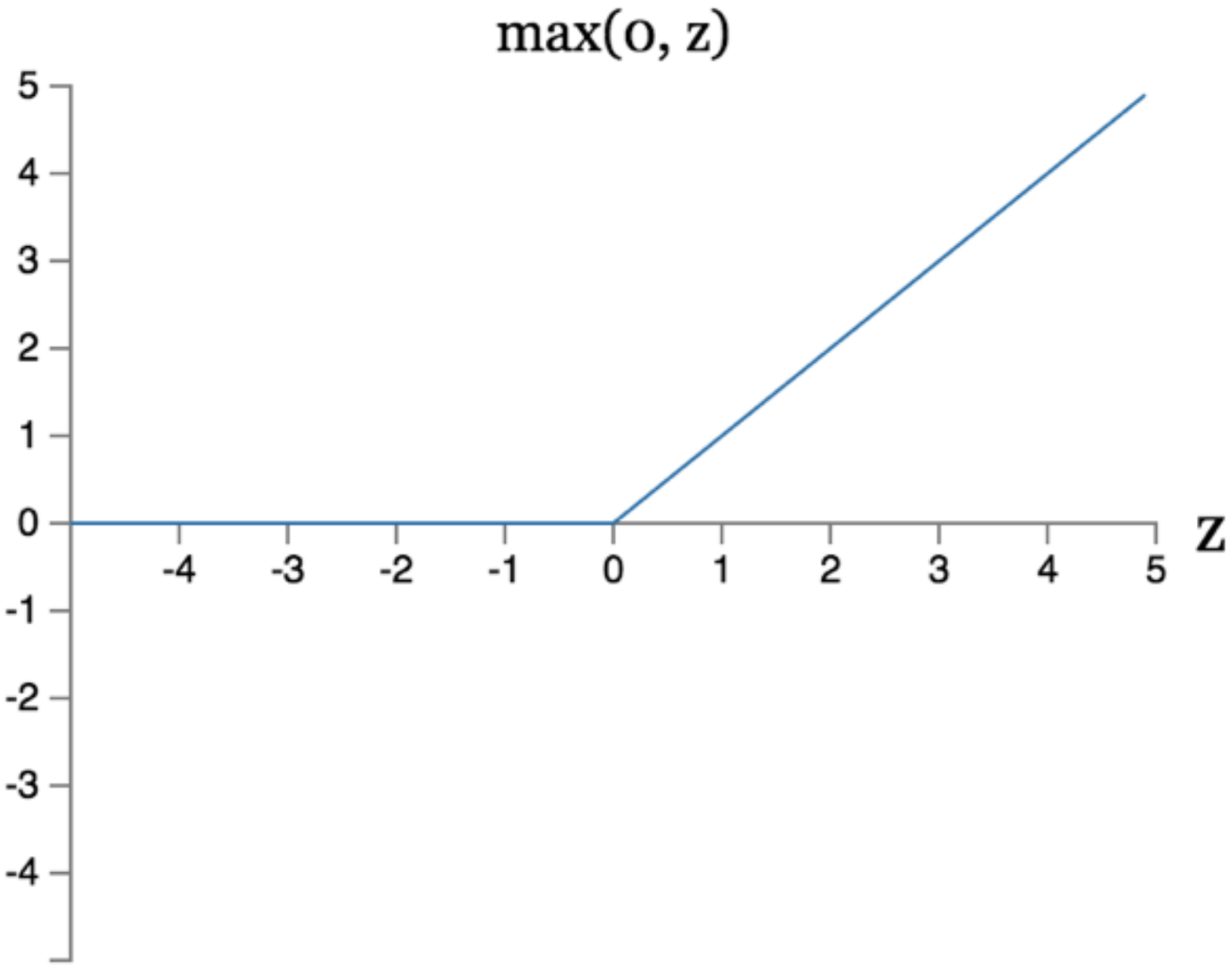
And why now?
GPU and datasets!

Deep Learning II

`Advanced` Neural Network techniques II

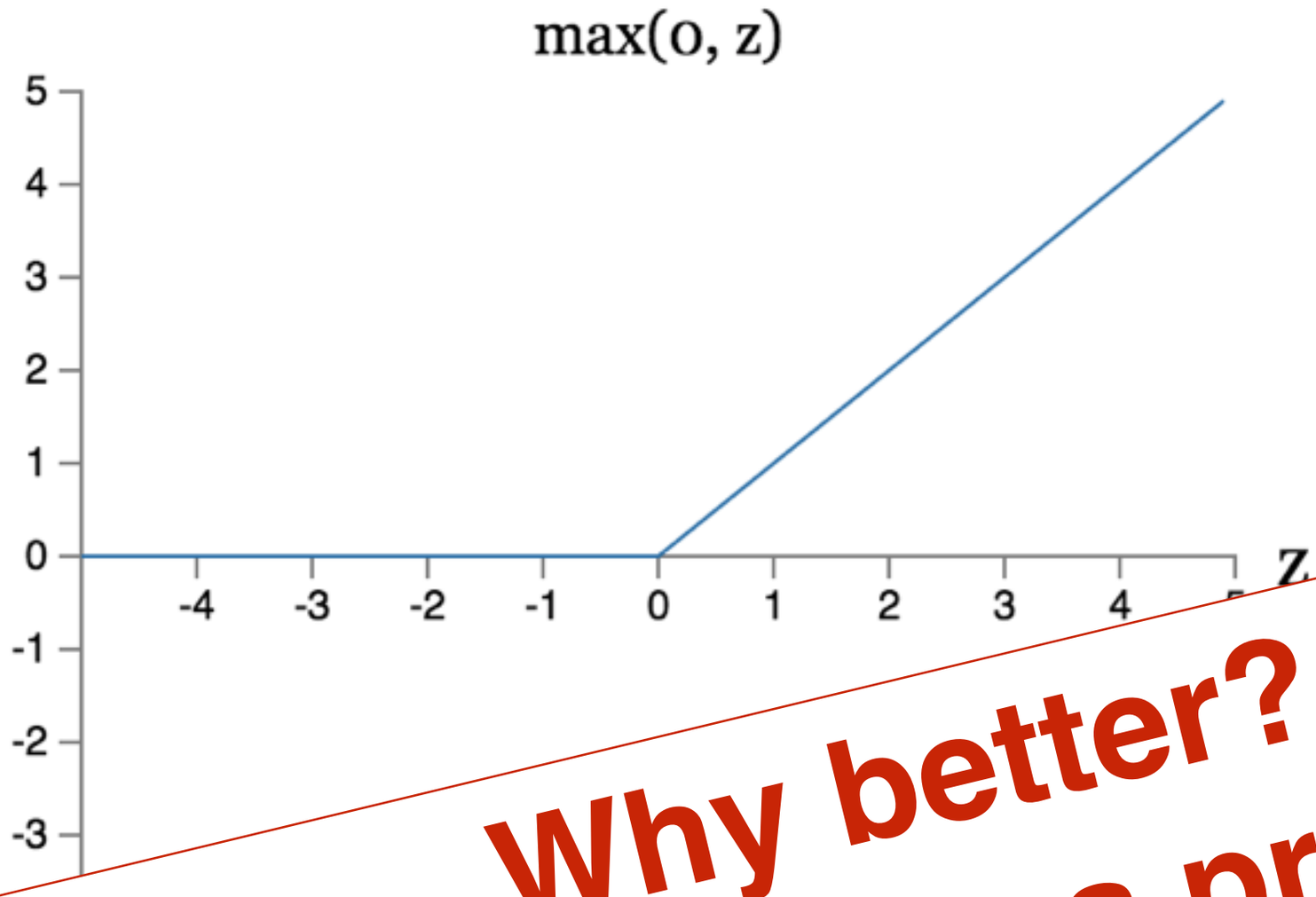
Rectified linear units

A better activation function



Rectified linear units

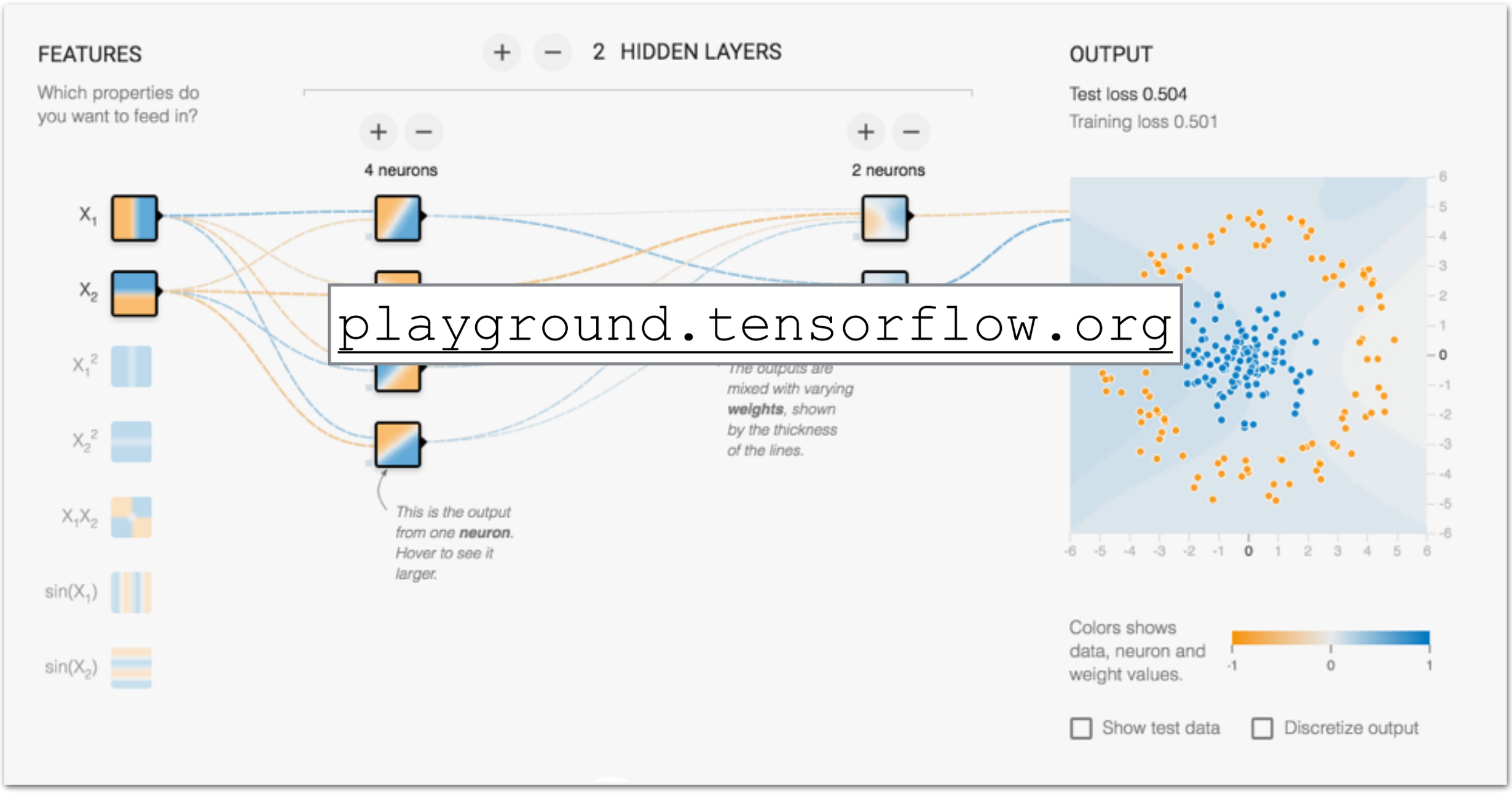
A better activation function



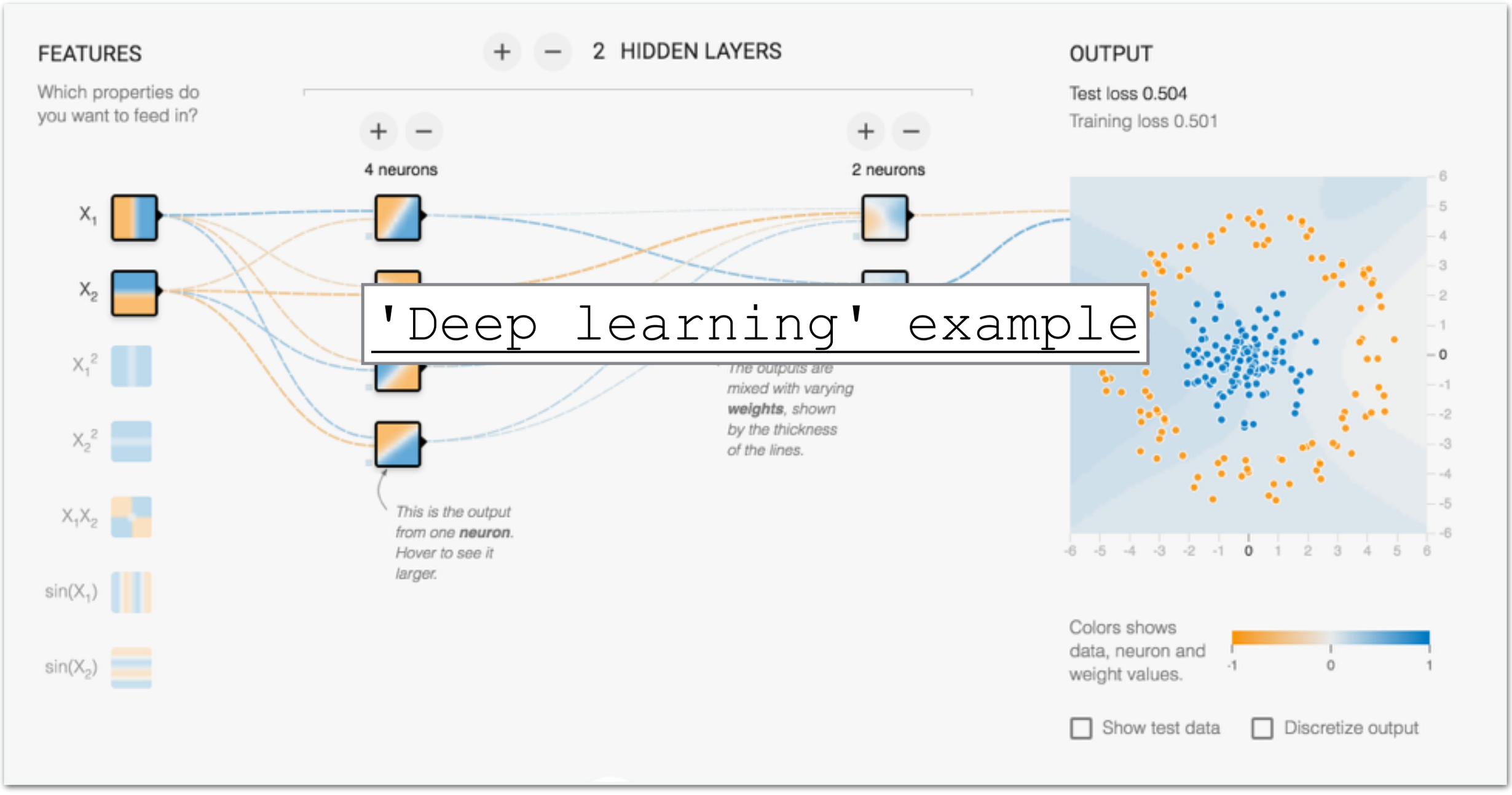
Why better?

What could be a problem?

Tensor flow Playground

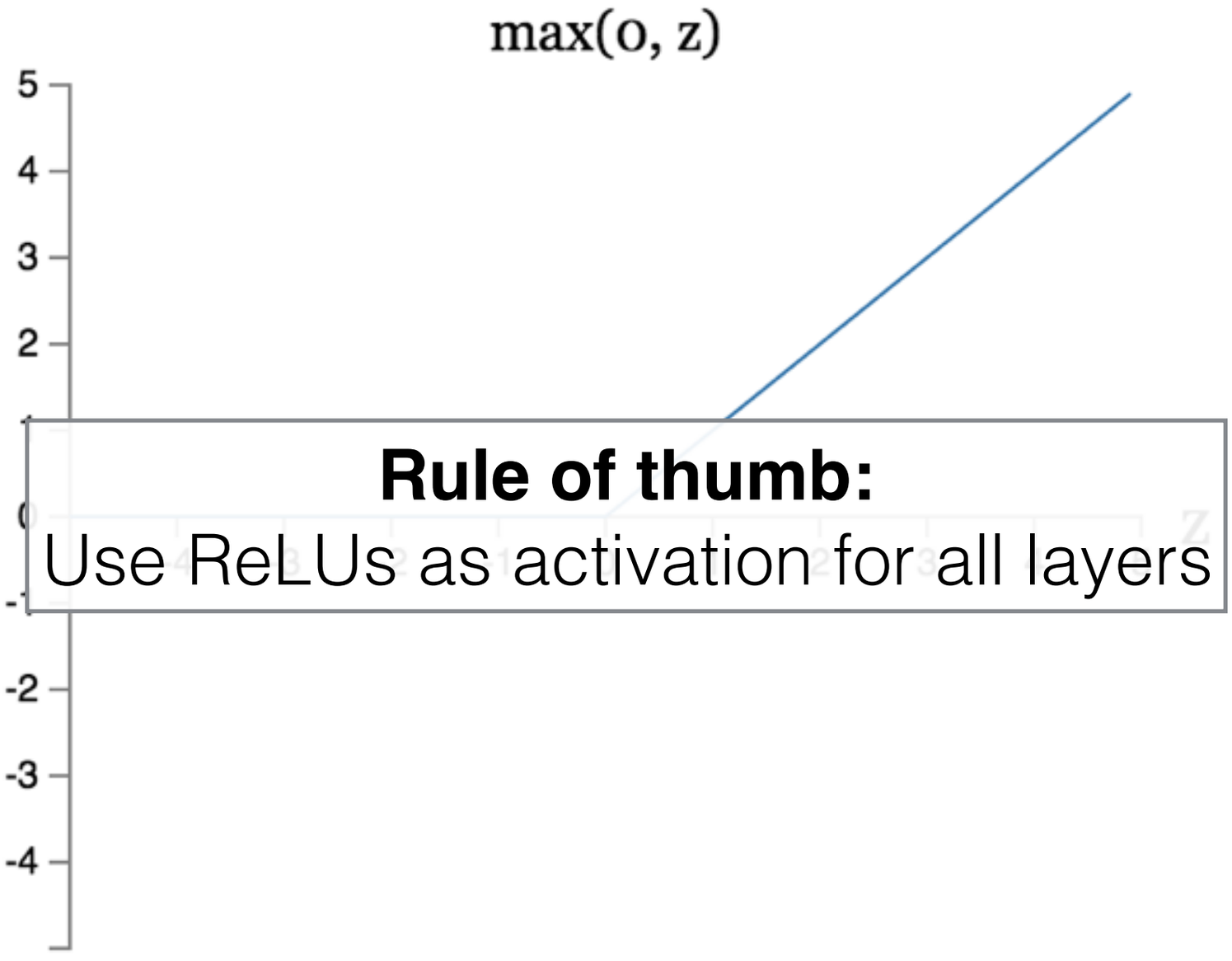


Tensor flow Playground



Rectified linear units

A better activation function



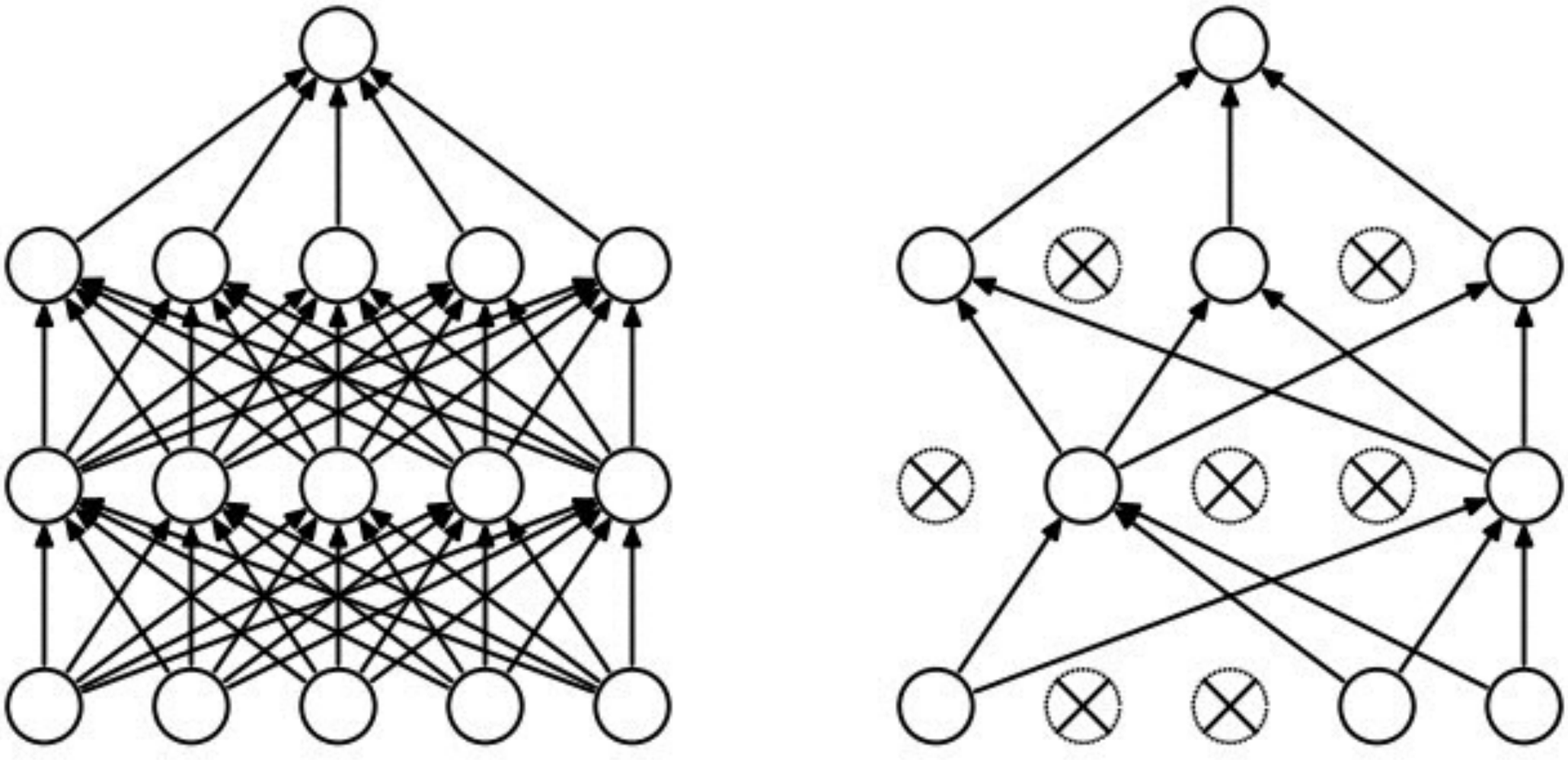
Rule of thumb:

Use ReLUs as activation for all layers

Why does Random Forests work so well?

Regularization with Dropout

random ensembles for NN



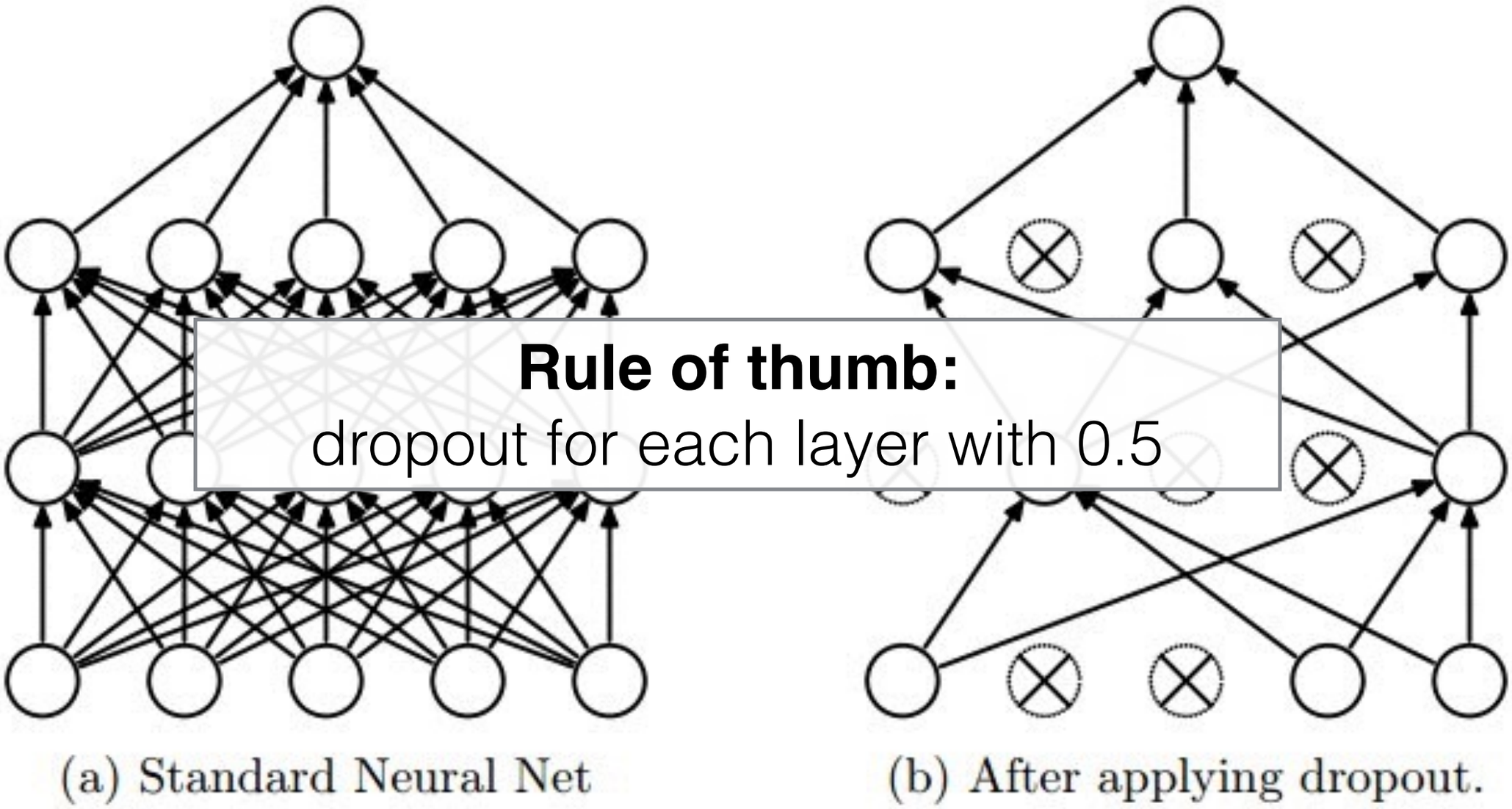
(a) Standard Neural Net

(b) After applying dropout.

Paper: “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”

Regularization with Dropout

random ensembles for NN



(a) Standard Neural Net

(b) After applying dropout.

Rule of thumb:
dropout for each layer with 0.5

Paper: “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”

Pimp my Gradient Descent

Decaying the learning rate

- Every n-epochs by a constant factor
- Every step exponentially
- When the learning stalls
- ….

Gradient Descent

$$w \rightarrow w' = w - \eta \nabla C$$

∇C = gradient of C
w = old weight
w'= new weight
η = learning rate (greek eta)

Gradient Descent and ravines

http://sebastianruder.com/optimizing-gradient-descent/

$$w \rightarrow w' = w - \eta \nabla C$$

∇C = gradient of C
w = old weight

w'= new weight
η = learning rate (greek eta)

Gradient Descent with Momentum

$$v \rightarrow v' = \mu v - \eta \nabla C$$

$$w \rightarrow w' = w + v'.$$

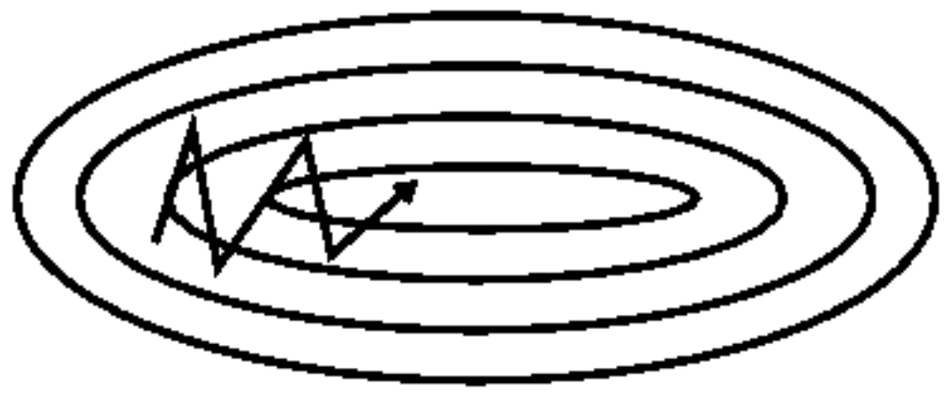
∇C = gradient of C
w = old weight
w'= new weight
η = learning rate (greek eta)

v = old velocity
v' = new velocity
μ = momentum co-efficent or friction (greek mu)

Gradient Descent with Momentum

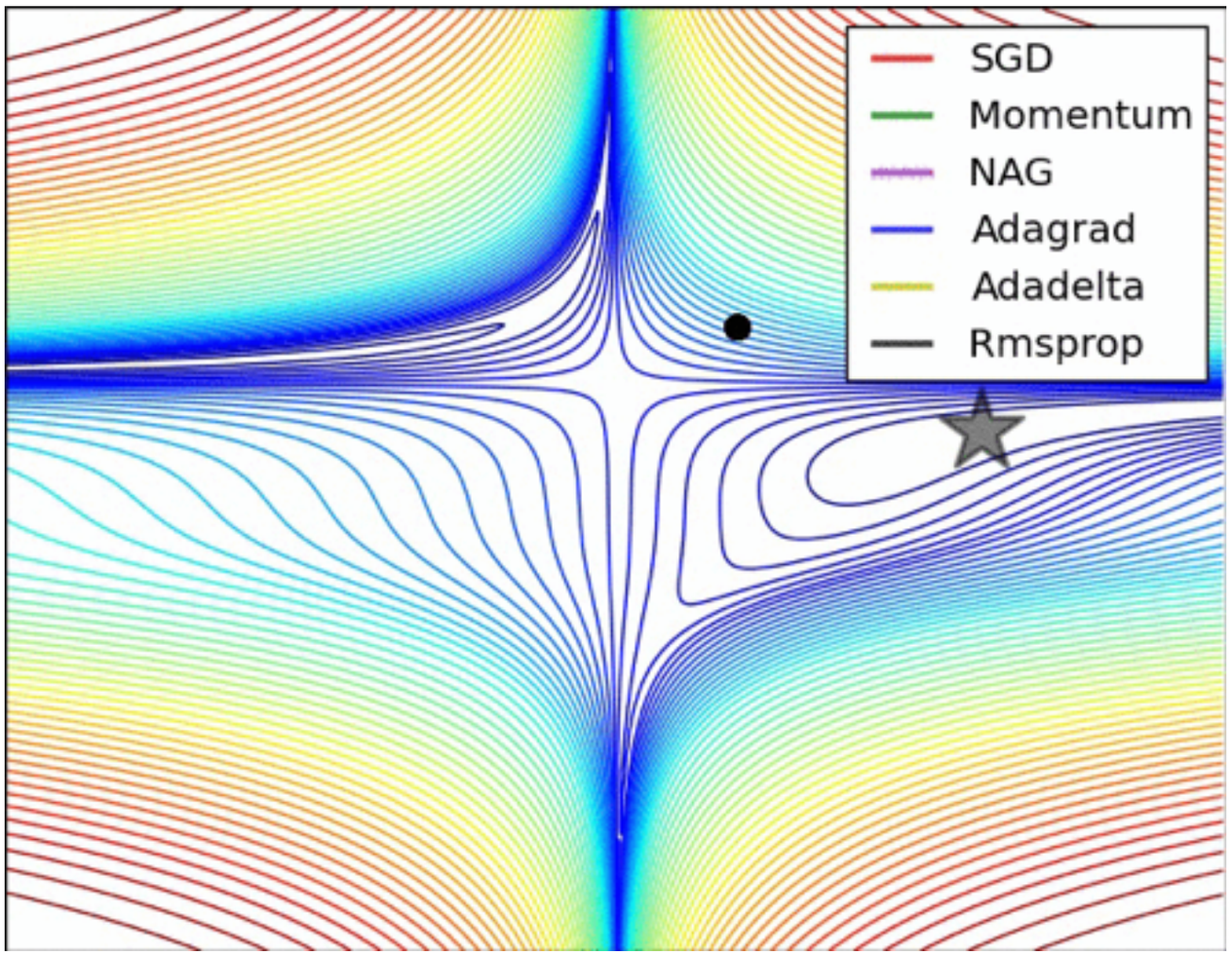
vanilla GD

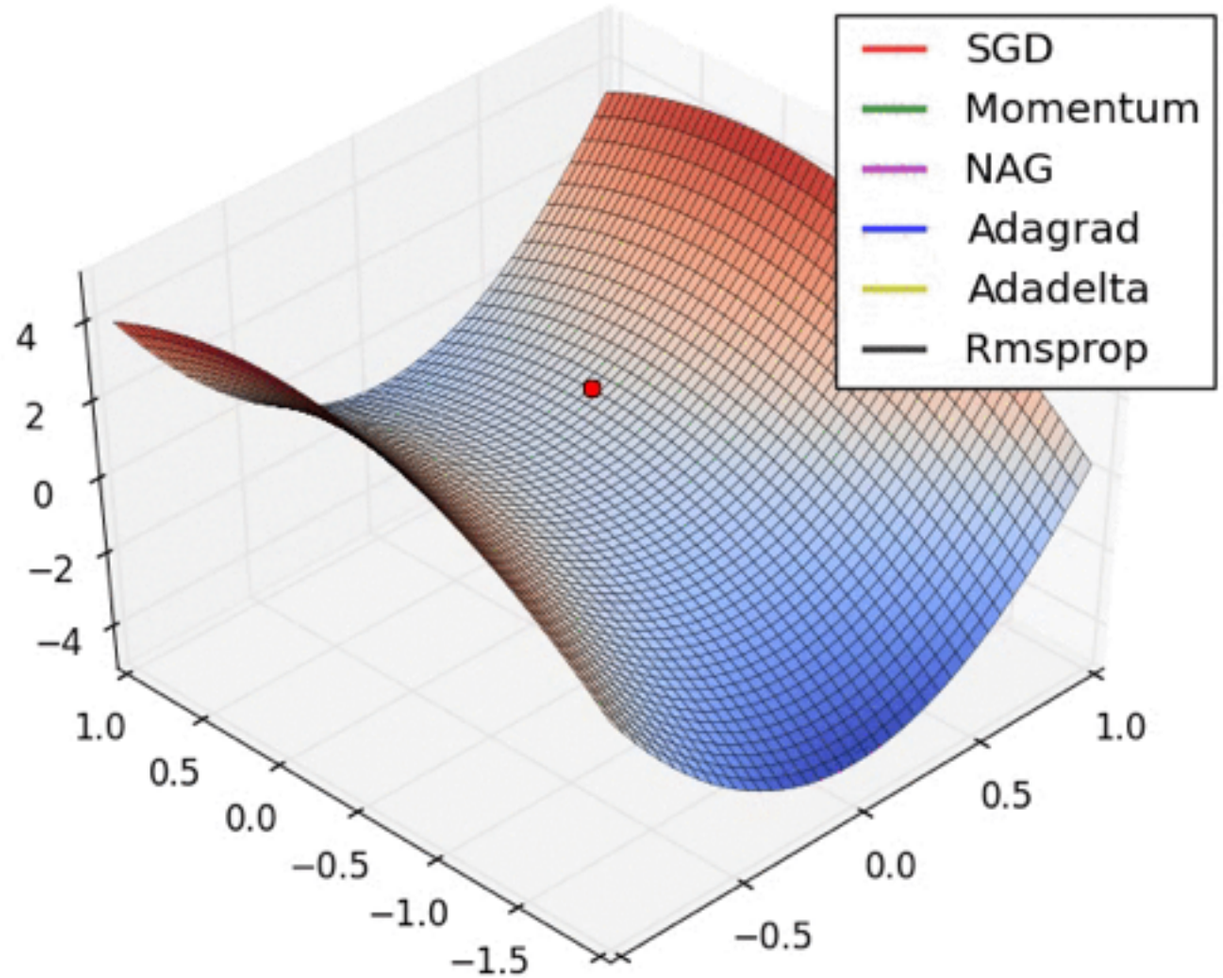
momentum SGD



http://sebastianruder.com/optimizing-gradient-descent/

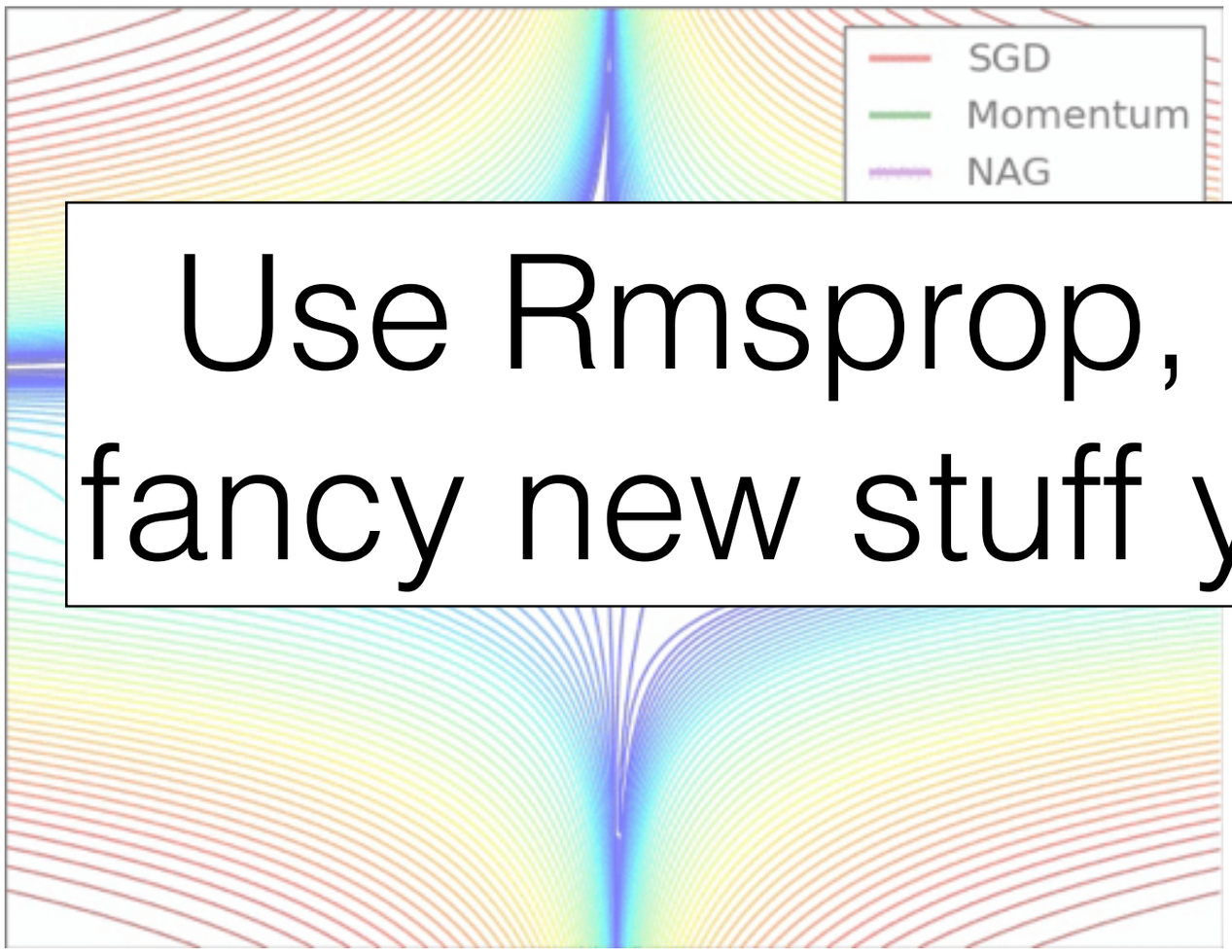
But there are so many others…

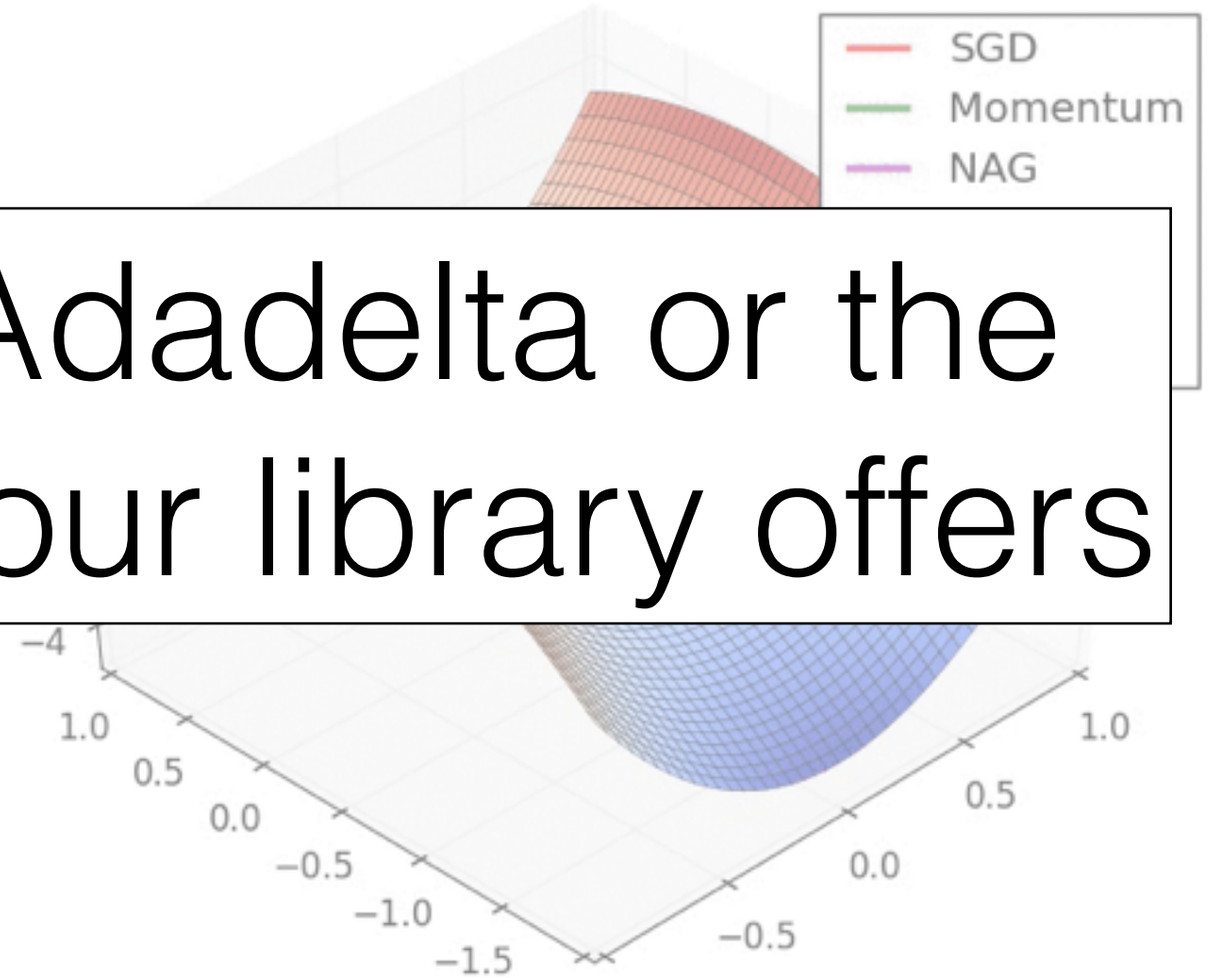




https://twitter.com/alecrad

But there are so many others…





Use Rmsprop, Adadelta or the fancy new stuff your library offers

https://twitter.com/alecrad

back to

PROGRAMMING

Try out :

- dropout
- ReLUs
- **Different optimizers**
- **https://keras.io/**

`~/keras_mnist.ipynb`

Deep Learning and data sets

MNIST (1998)

- Hand written digits (0-9)
- Collected by the United States' National Institute of Standards and Technology, NIST
- Pre segmented
- 28x28 pixel in size
- 60k train, 10k test
- ('M' stands for modified)

MNIST 1998

Eric Jonas
@stochastician

Follow

my advisor: “MNIST is a unit test”

9:39 PM - 25 Jan 2016

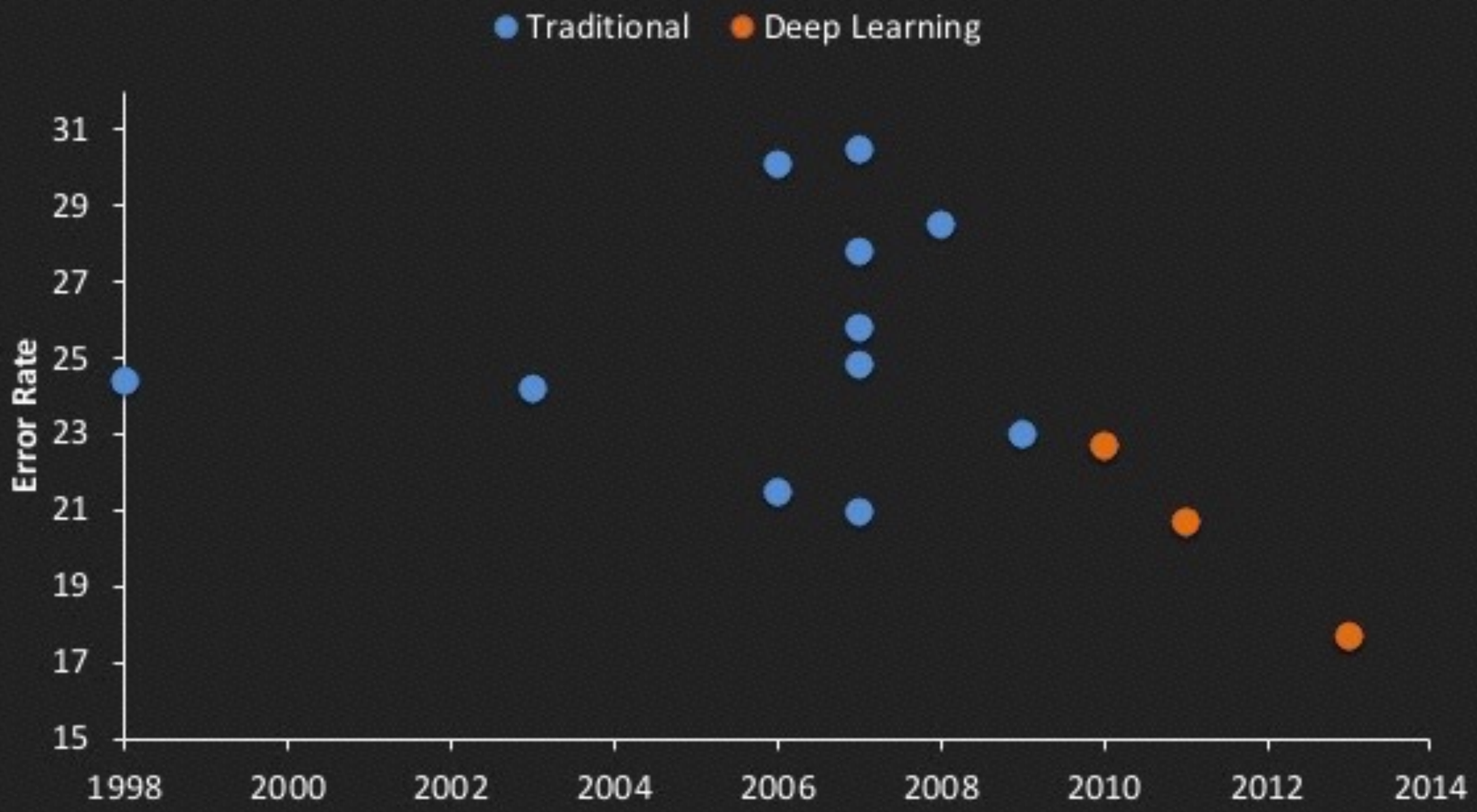
7 11

- Hand written digits (0-9)
- Collected by the United States' National Institute of Standards and Technology, NIST
- Pre segmented
- 28x28 pixel in size
- 60k train, 10k test
- ('M' stands for modified)

MNIST 1998

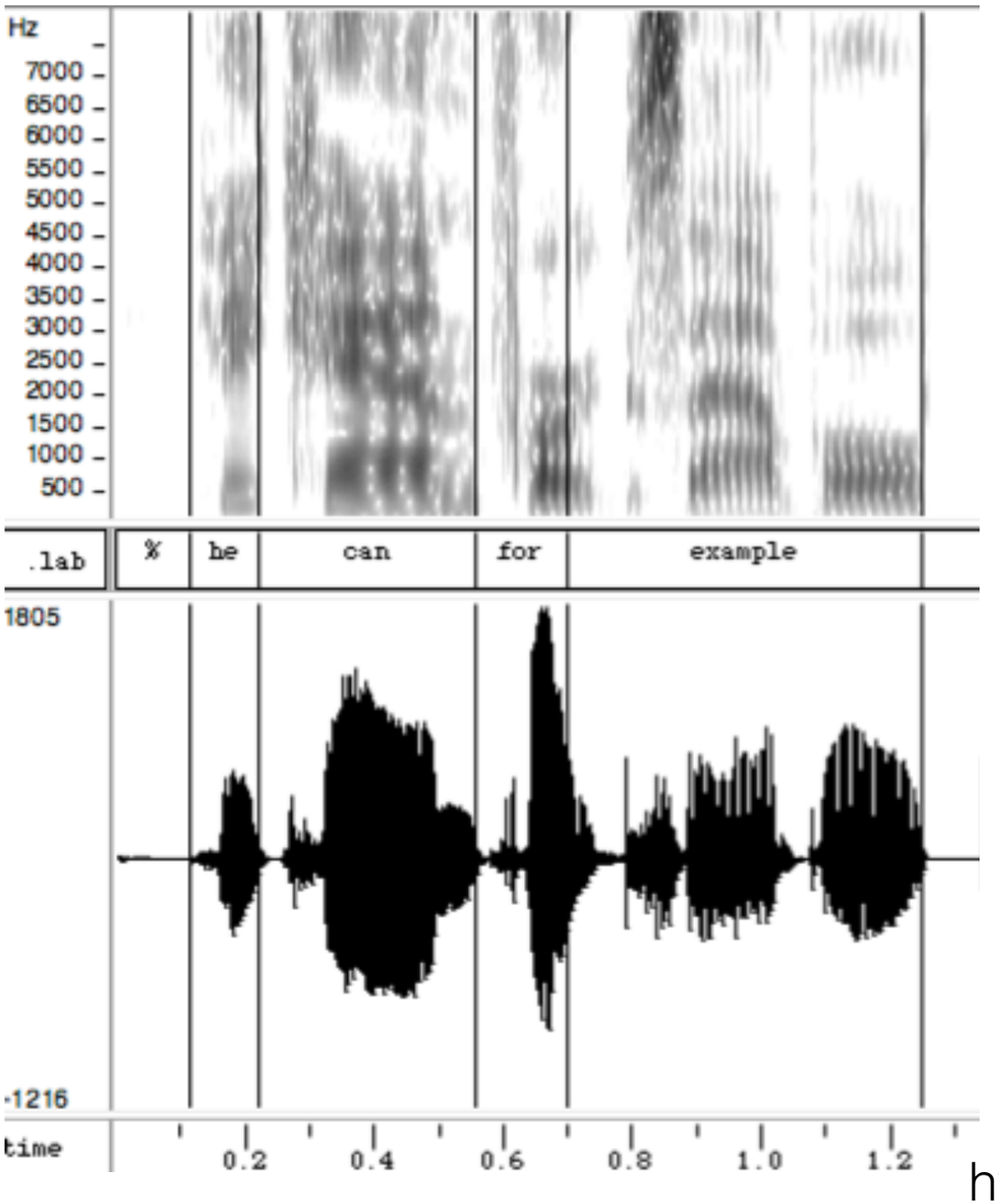
- Hand written digits (0-9)
- by the United
  tional Institute of
  and Technology,
- segmented
- 28x28 pixel in size
- 60k train, 10k test
- ('M' stands for modified)

TIMIT Speech Recognition



clarifai

TIMIT (1990)



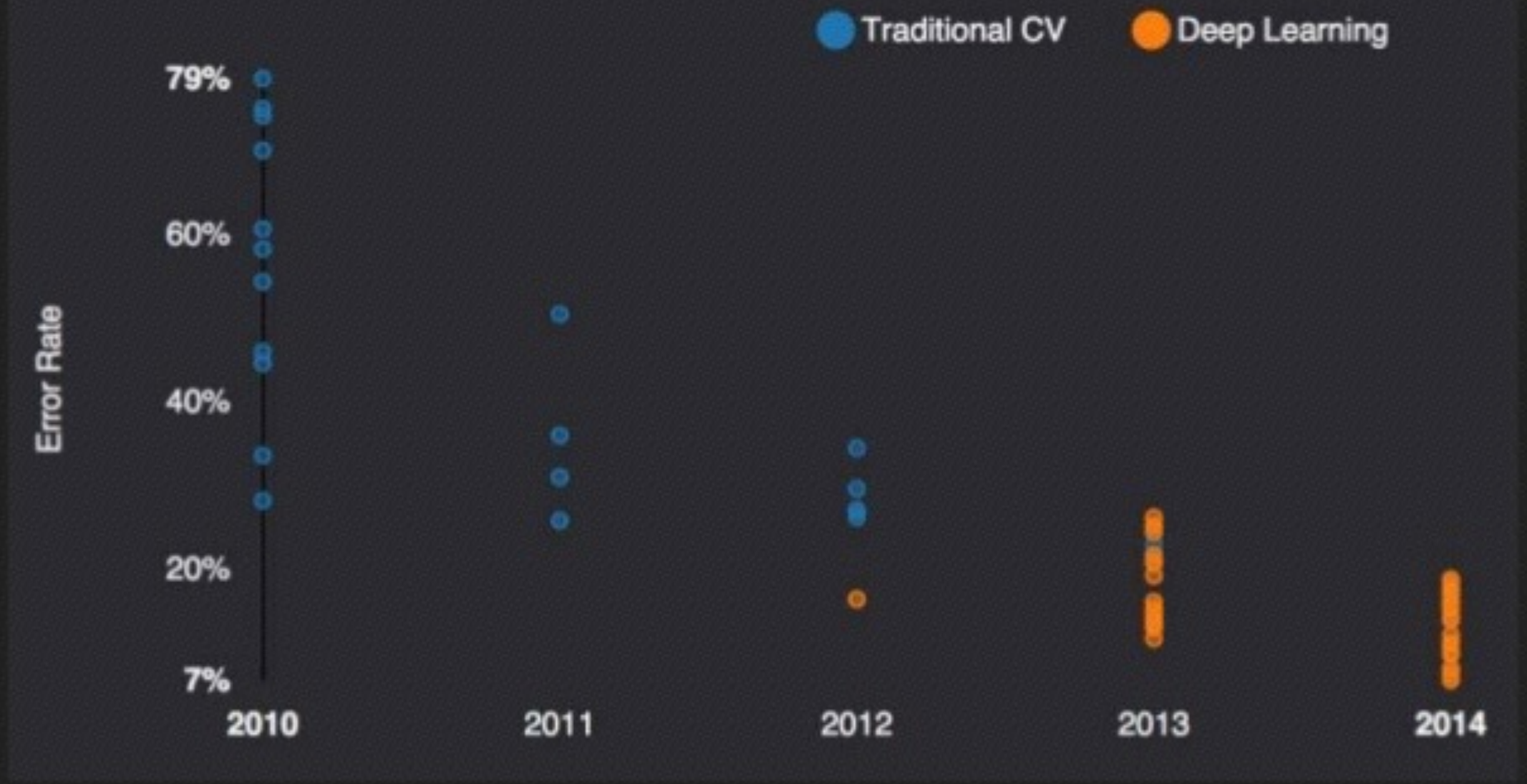
- Texas Instruments (TI) and Massachusetts Institute of Technology (MIT)
- 10 english sentences
- 630 speakers
- tagged on phone level
- 61 different phones
- 5.4h material
- @16 kHz with 16 bit

➡ (~400m samples)

➡ https://catalog.ldc.upenn.edu/LDC93S1

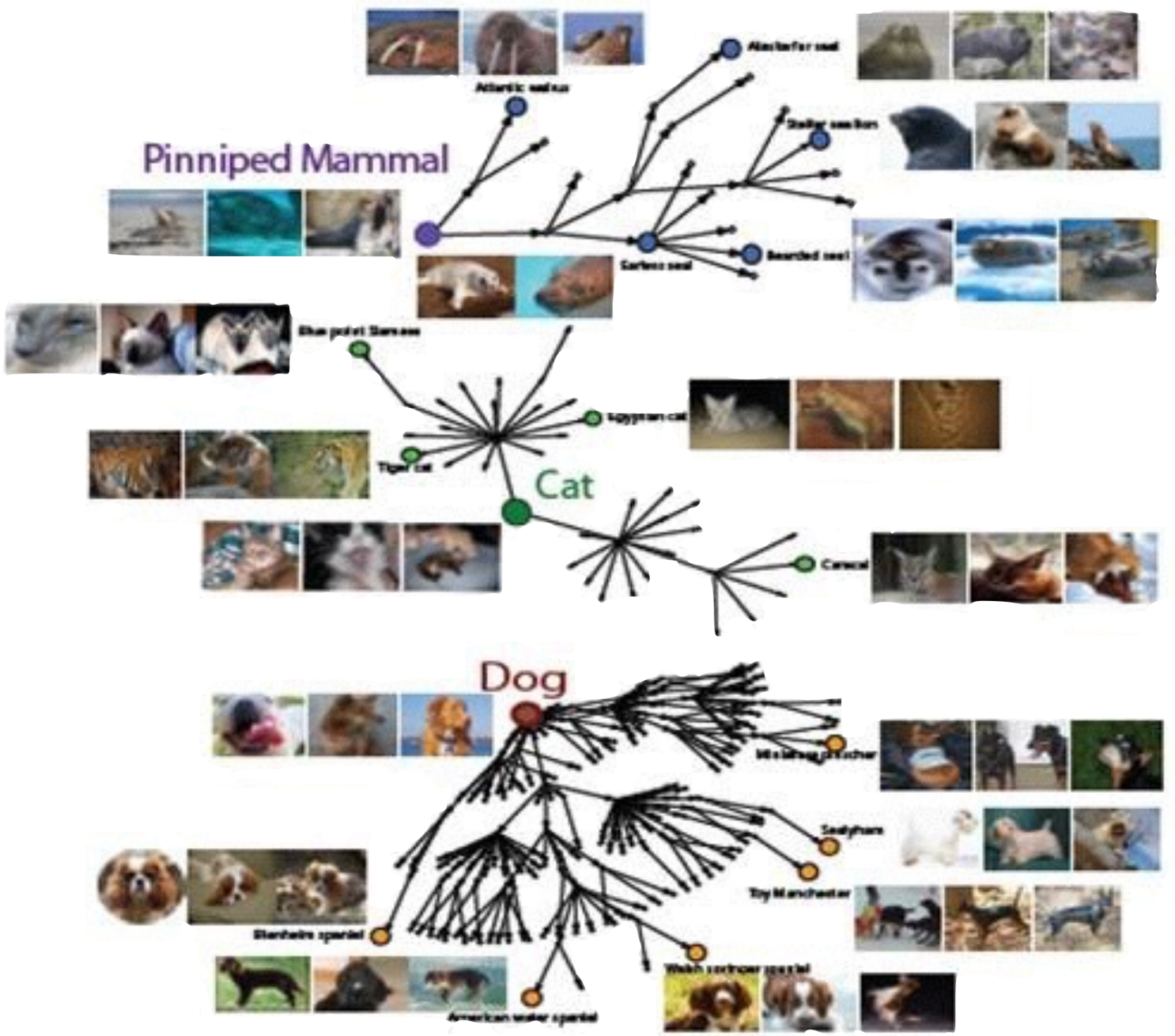
http://languagelog.ldc.upenn.edu/nll/?p=8116

ImageNet Error Rate 2010-2014



clarifai

‘Imagenet’ (2012)



Precise name: **Large Scale Visual Recognition Challenge**

- 1m images
- 1k concepts
- subset of the Imagenet
  - 20m images
  - 10k concept (tree structure)
  - still growing
  - image-net.org

http://web.eecs.umich.edu/~jiadeng/

‘MS COCO’ (2014)

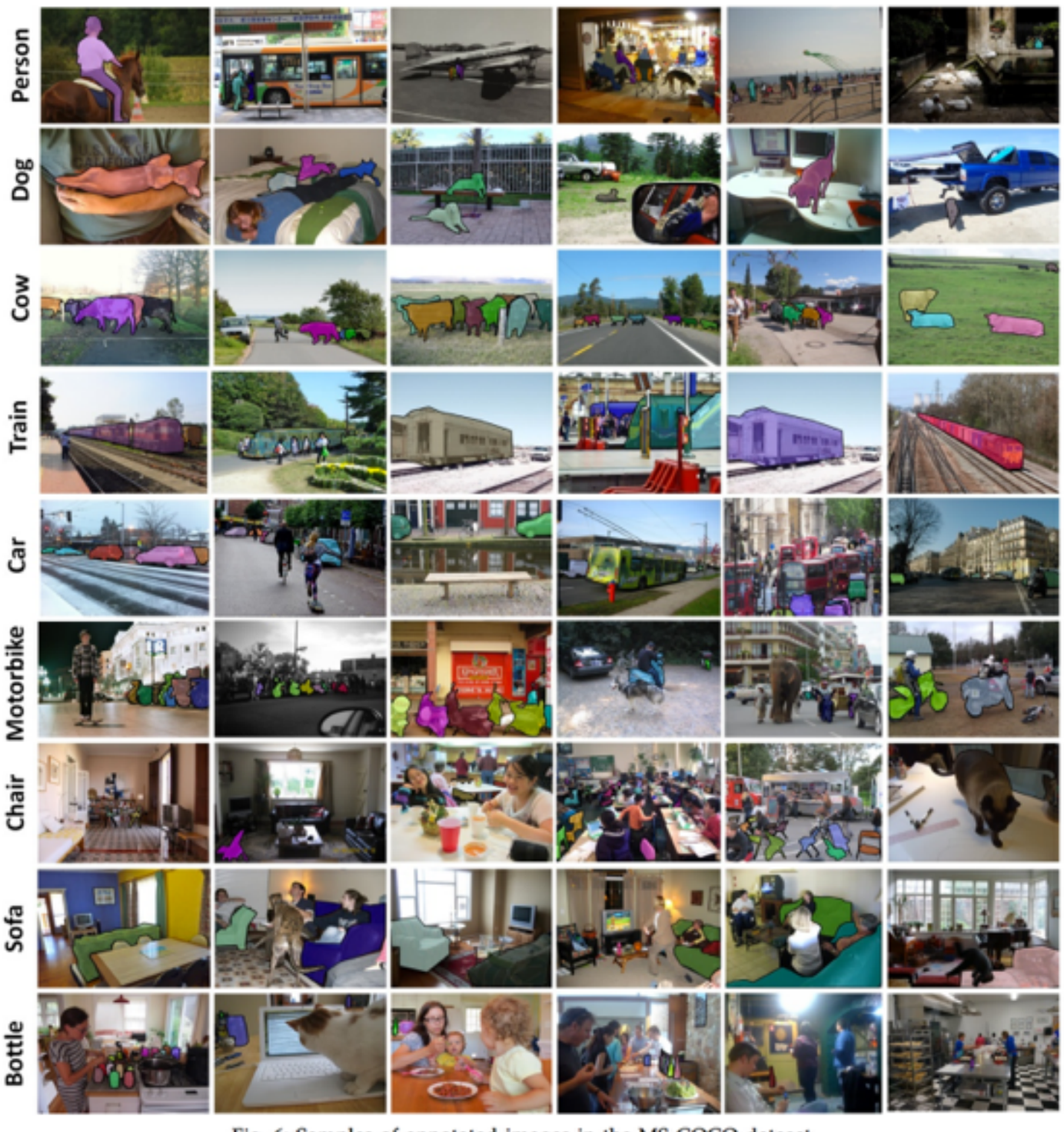
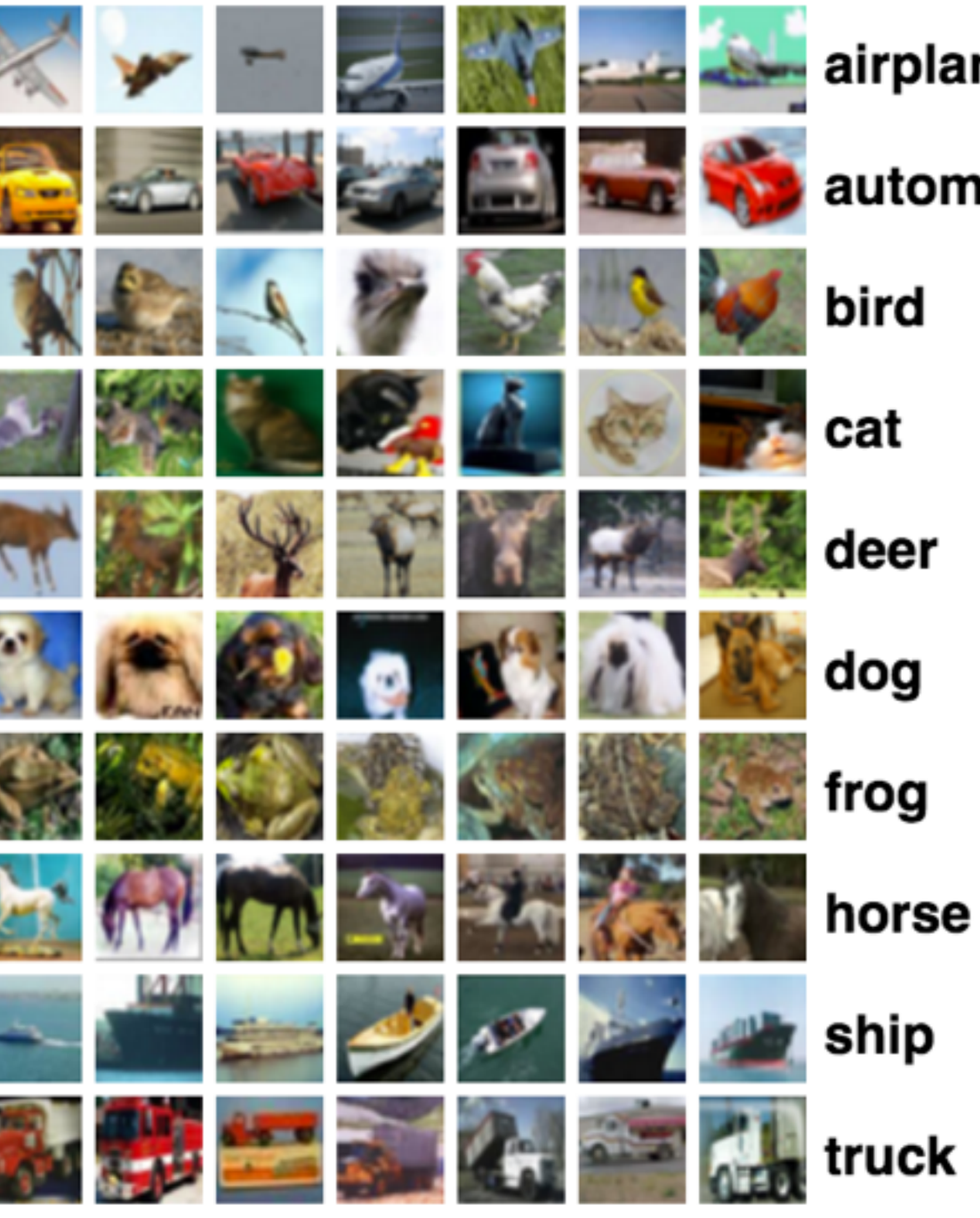


Fig. 6: Samples of annotated images in the MS COCO dataset.

Precise name: **Microsoft’s Common Objects in COntext**

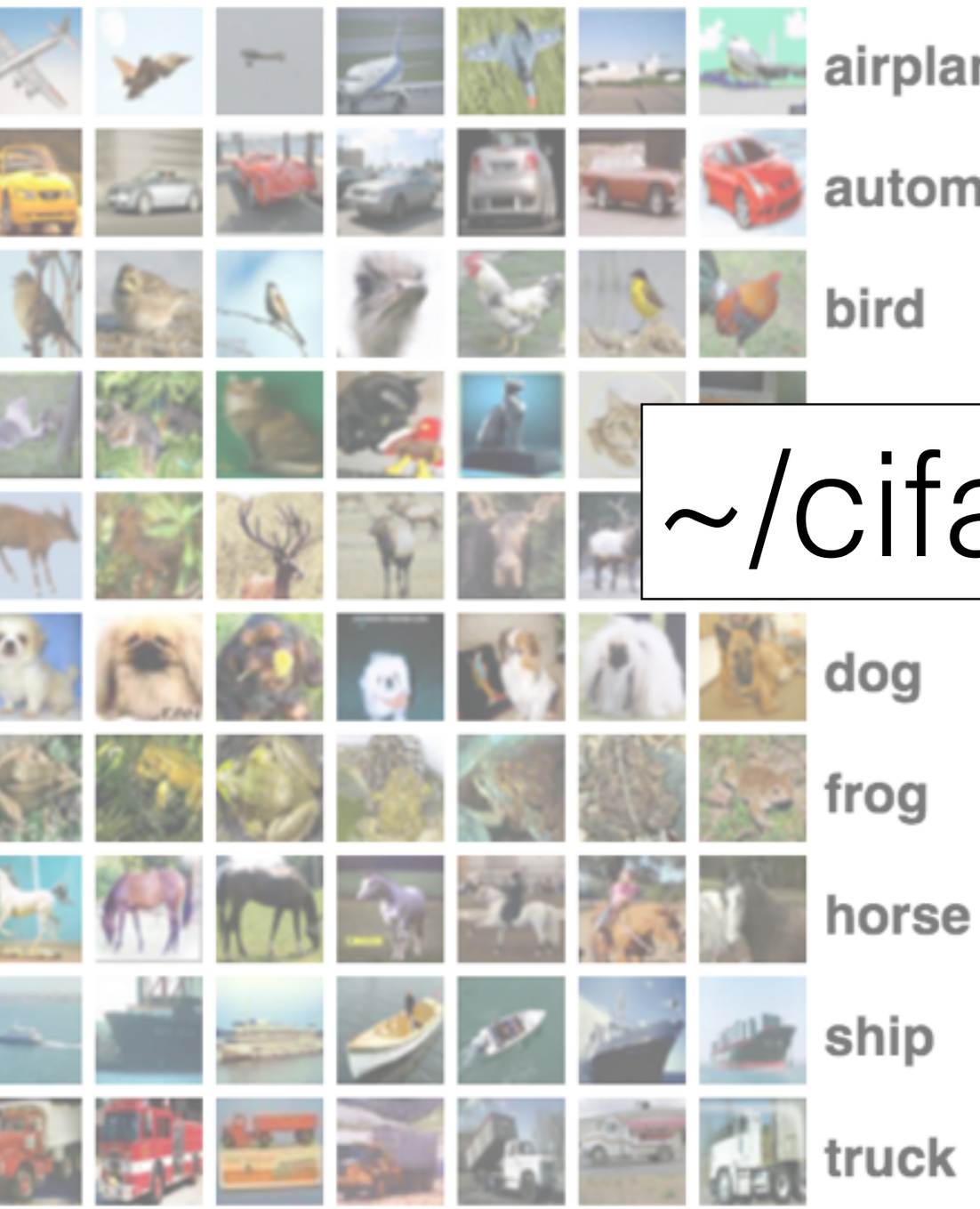
- More than 70 categories
- Object segmentation
- Multiple objects per image
- More than 300k images (150k)
- **5 captions per image**
- mscoco.org

CIFAR-10



- CIFAR stands for C*anadian Institute for Advanced Research*
- 10 concepts
- 6000 images per concept
- Image size is 32x32x3
- Part of the 80 Million Tiny Images dataset
- https://www.cs.toronto.edu/~kriz/cifar.html

CIFAR-10



~/cifar10.ipynb

- CIFAR stands for C*anadian Institute for Advanced Research*
- 10 concepts
- 6000 images per concept
- 32x32x3
- Part of the 80 Million Tiny Images dataset
- https://www.cs.toronto.edu/~kriz/cifar.html

Cat or dog?

Or how to work with little data

Cat or dog?

Or how to work with little data

~/cats+dogs/train/cats/…

from: www.kaggle.com/c/dogs-vs-cats

Cat or dog?

Or how to work with little data

~/cats+dogs/first_try.ipynb

from: www.kaggle.com/c/dogs-vs-cats

The trick

The ‘trick’

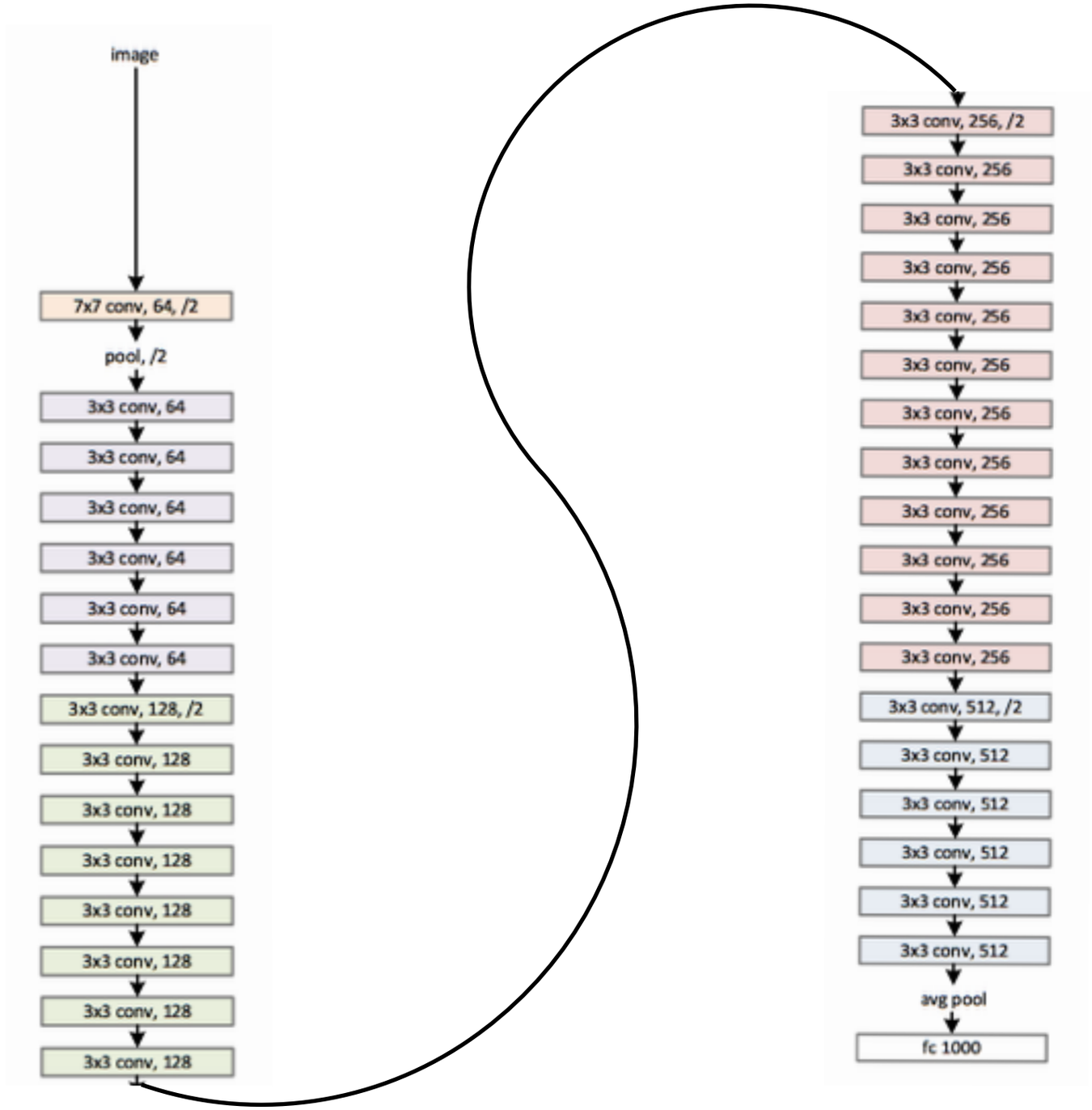


image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
avg pool
fc 1000

The ‘trick’

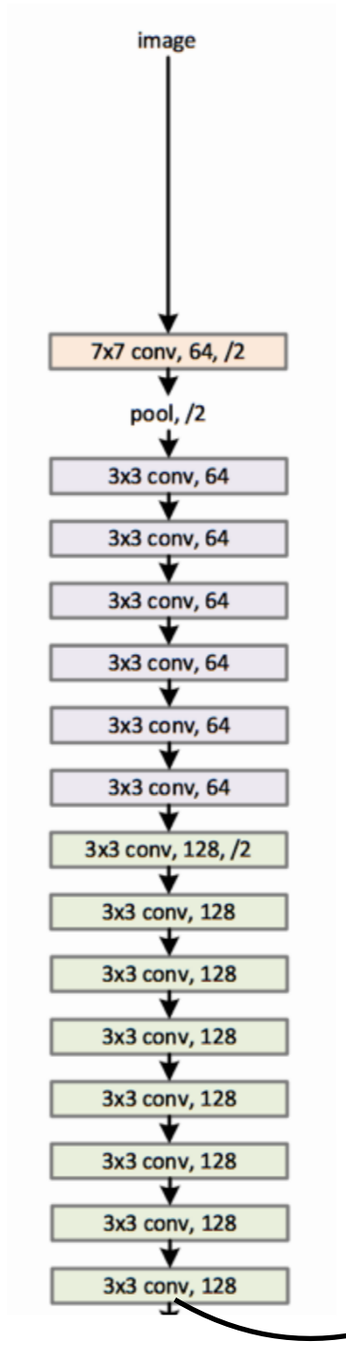
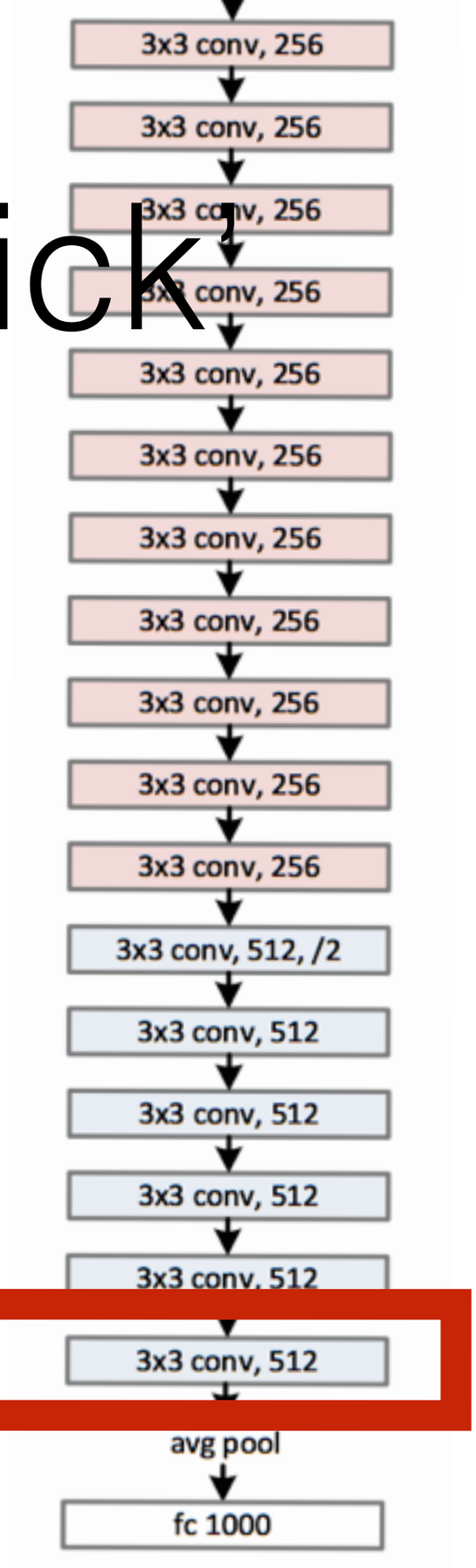
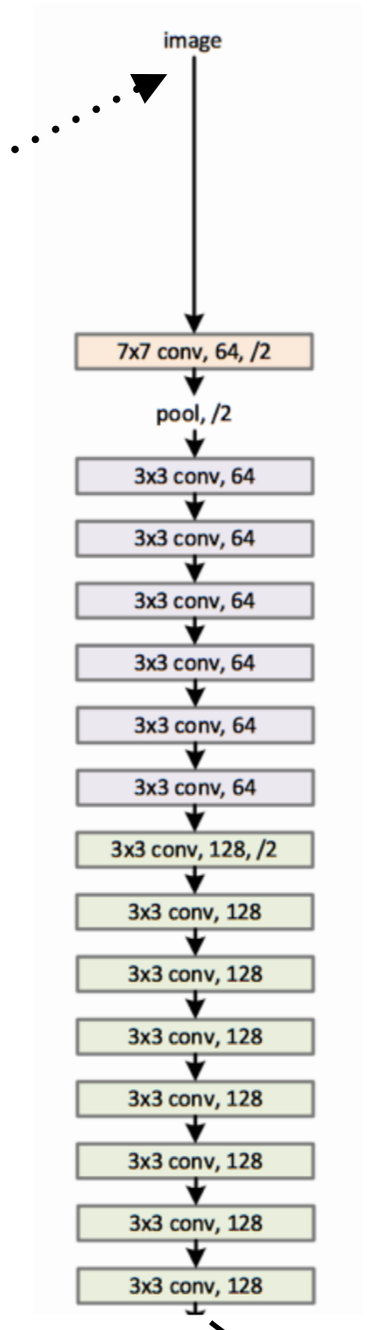


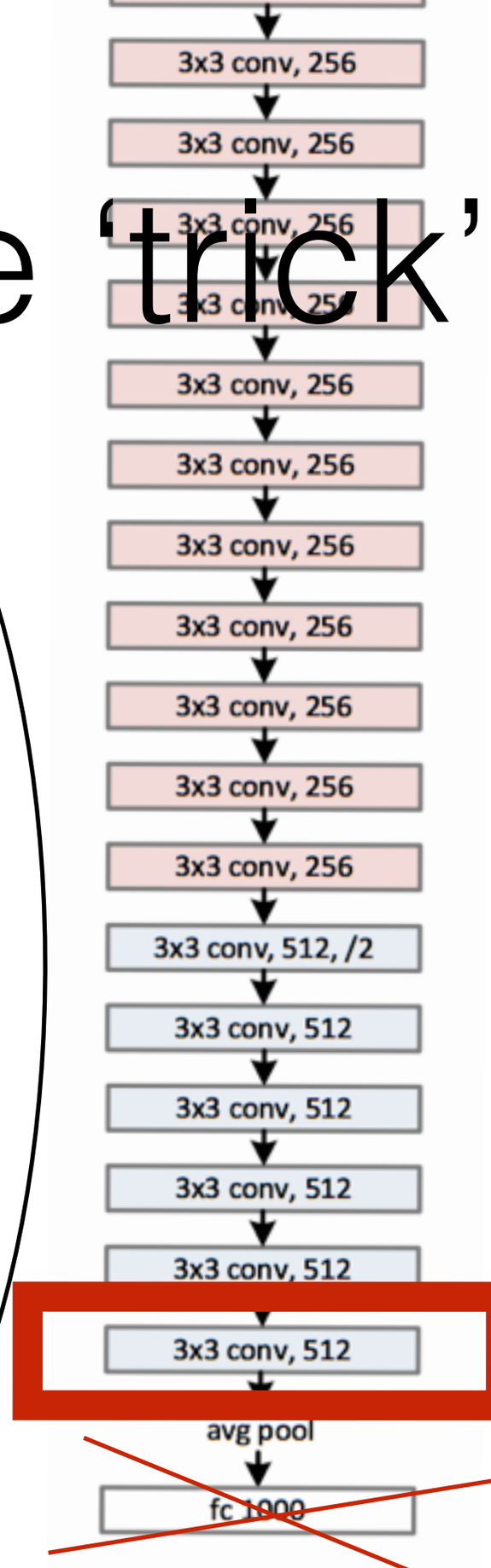
image
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128
3x3 conv, 128


3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
3x3 conv, 512
avg pool
fc 1000

The ‘trick’

images

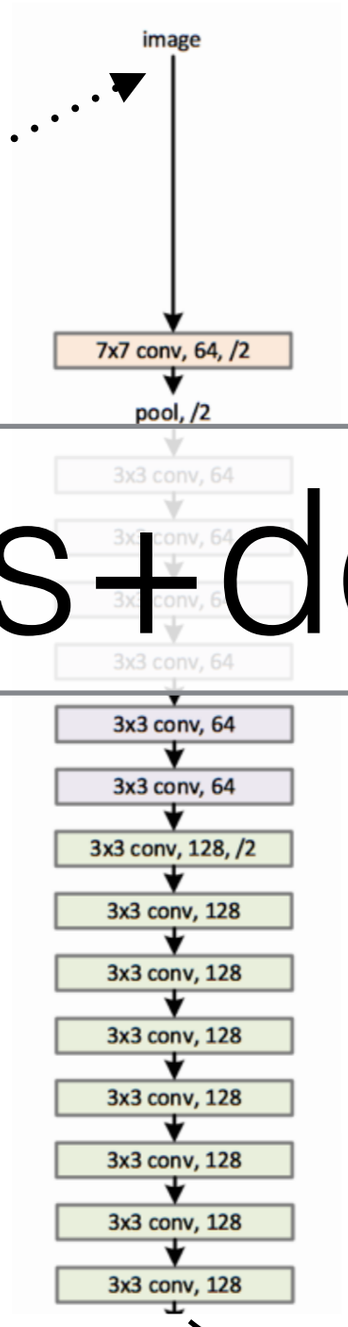


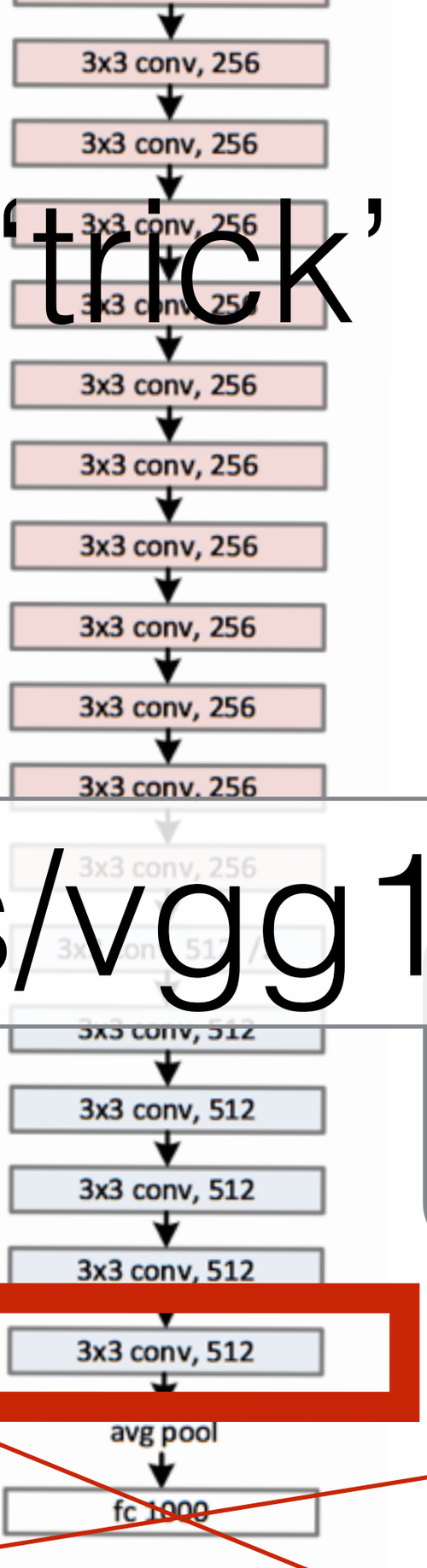


predictions

SVM | RF | NN| …

The ‘trick’



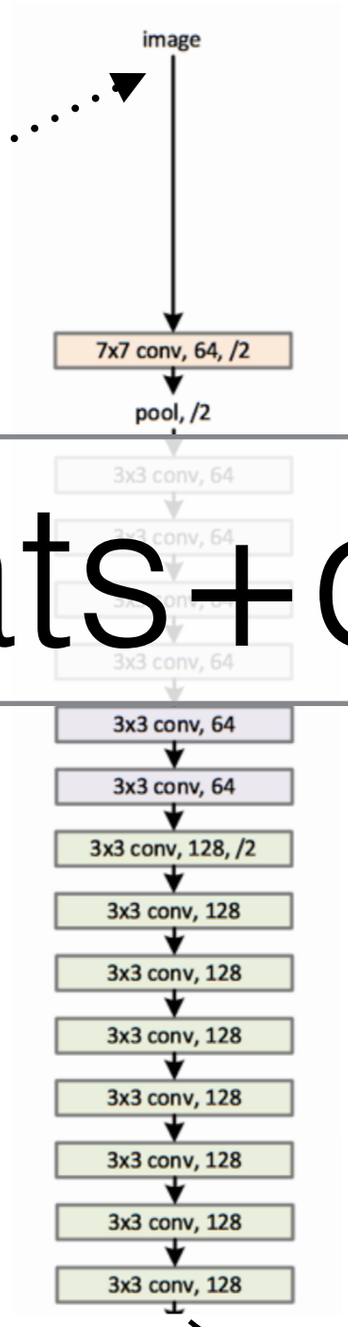


predictions

SVM | RF | NN| …

~/cats+dogs/vgg16.ipynb

The ‘trick’



predictions

SVM | RF | NN| …

~/cats+dogs/trick.ipynb

The ‘trick’

Using pre-trained weights of as feature extractor in combination with an familiar classifier like SVM or RF.

Pros

- predictable training time
- learned features
- needs less data

Cons

- needs pre-trained network weights
- introduction of bias
- reaches not the full potential of Deep Learning

The ‘trick’

Using pre-trained weights of as feature extractor in combination with an familiar classifier like SVM or RF.

Pros

- predictable training time
- learned features
- needs less data

Cons

- needs pre-trained network weights
- introduction of bias
- reaches not the full potential of Deep Learning

~/cats+dogs/trickier.ipynb

Q: How do I know what architecture to use?

Q: How do I know what architecture to use?

A: Don’t be a hero:

1. Take what ever works on ILSVRC (latest ResNet)
2. Download a pretrained model
3. Potentially add/delete some parts of it
4. Finetune to your application

Karpathy at BayArea DL School 2016

Q: How do I know what architecture to use?

A: Don’t be a hero:

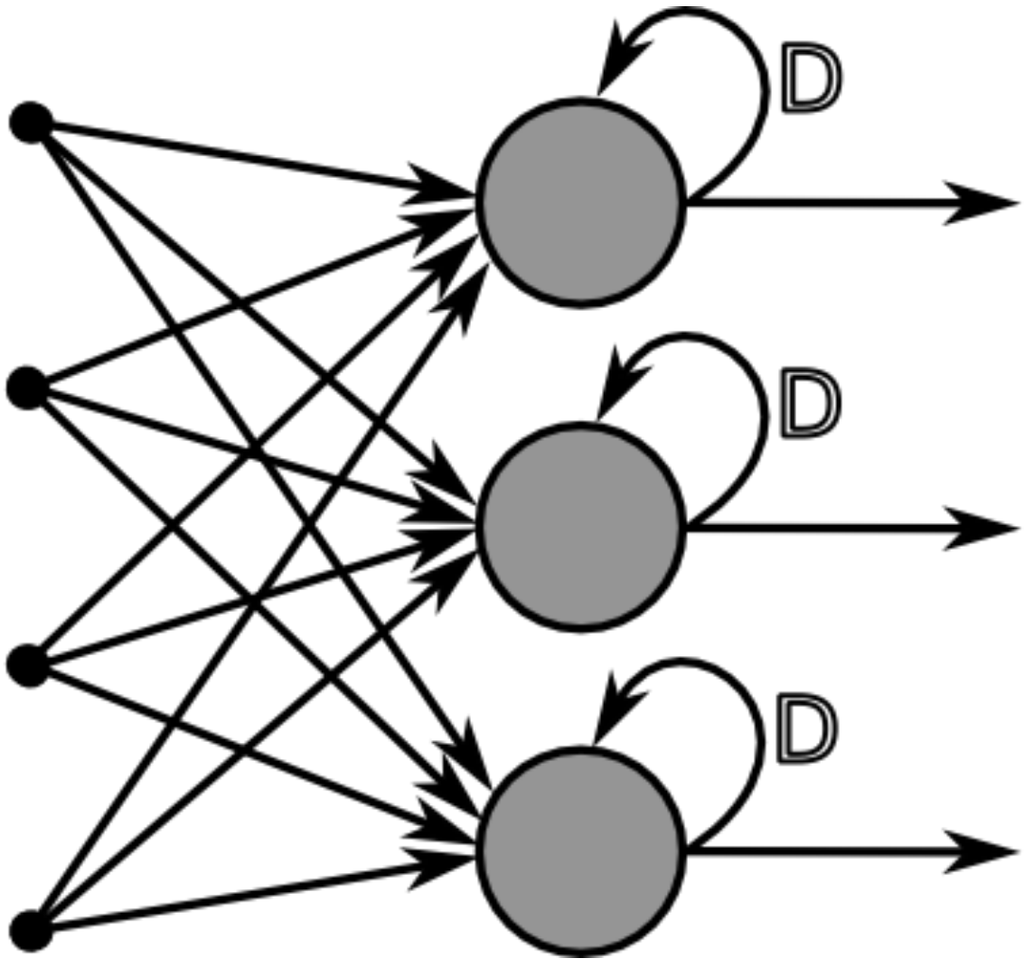
1. Take what ever works on ILSVRC (latest ResNet)
2. Download a pretrained model
3. Potentially add/delete some parts of it
4. Finetune to your application

Architecture engineering
is the new *Feature engineering*

Karpathy at BayArea DL School 2016

Other Deep Learning approaches

Recurrent Neural Networks (RNN)



- Good for time series problems
  - sound
  - hand writing
- unstable
  - vanishing and exploding gradients
- but recent advancement
  - with major breakthroughs in voice recognition (google now, siri, cortana)

https://commons.wikimedia.org/wiki/File:RecurrentLayerNeuralNetwork.png

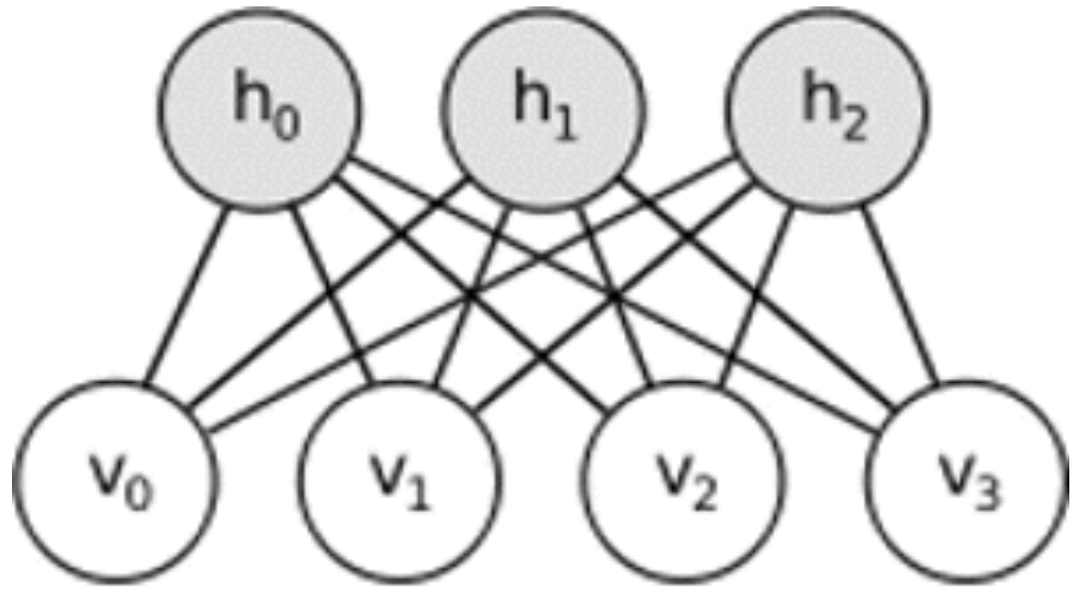
Restricted Boltzman Machines (RMBs)

- are also know as **Deep Believe Networks**

- train a reconstruction of the inputs with fewer parameters

- can be trained unsupervised or semi-supervised

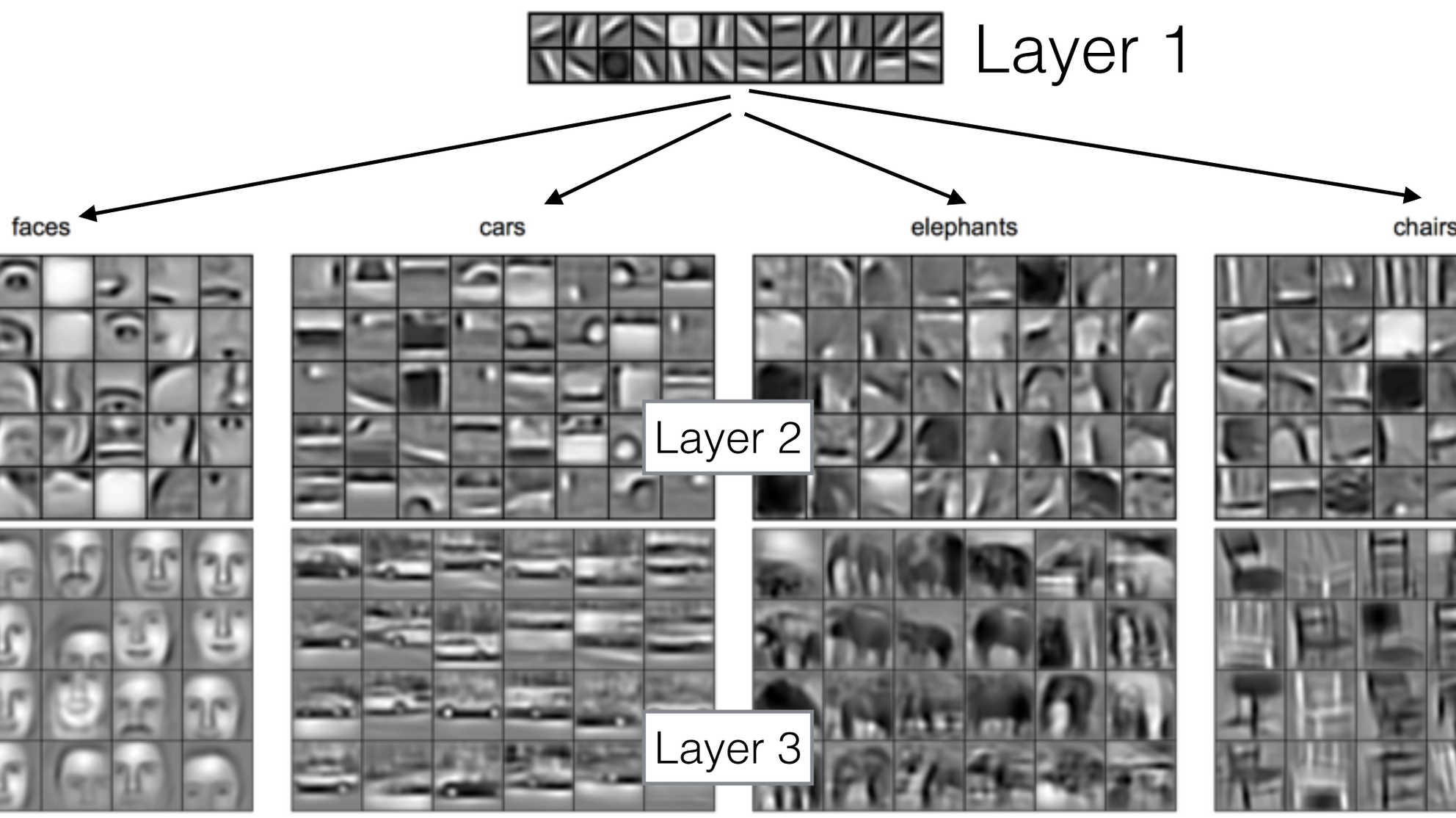
- were the first general-purpose deep architecture that could be trained successfully (2006)



- give easily interpretable nodes

- are now surpassed by CNNs

Convolutional Deep Believe Networks



Paper: Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations

Deep Learning small talk

The Deep Learning Heads

Yan LeCun

- Known for CNNs
- LeNet5 1989 (AT&T)
- Torch / Lua
- now head of Facebook AI Research (working there also with V. Vapnik (SVM))

Geoffrey Hinton

- Co-inventor of the backpropagation in ’74
- Known for DBNs
- University Toronto + 50% google research since 2012
- Vision break through 2012

Yoshua Bengio

Yoshua Bengio

- University Montreal
- Sep 2016: Grant of $93 million to work on "AI"
- Theano

Jürgen Schmidhuber

- Pronounciation:
  You_again Shmidhoobuh
- Known for RNN
- University of Lugano (Switzerland)
- Also working on art theory + generation through NNs

The Neural Computation and Adaptive Perception program

- founded by LeCun, Hinton and Bengio in 2004

- funded by a Canadian Institute for Advanced Research (CIFAR) grant

- invite only

- laid the groundwork and spearheaded the Neural Network renaissance

The Neural Computation and Adaptive Perception program

?

Deep Learning Conspiracy

http://people.idsia.ch/~juergen/deep-learning-conspiracy.html

OpenAI

- non-profit artificial intelligence research company
- founded by 2015 by Elon Musk and Sam Altman (YCombinator)
- ‘1 Billion $ in funding’

“*Carefully promote and develop friendly AI in such a way as to benefit, rather than harm, humanity as a whole.”*

OpenAI

- non-profit artificial intelligence research company

OpenAI's Director of 'Strategy and Communications', Jack Clark, has a very good newsletter about Deep Learning and AI.

Follow it to stay up to date:

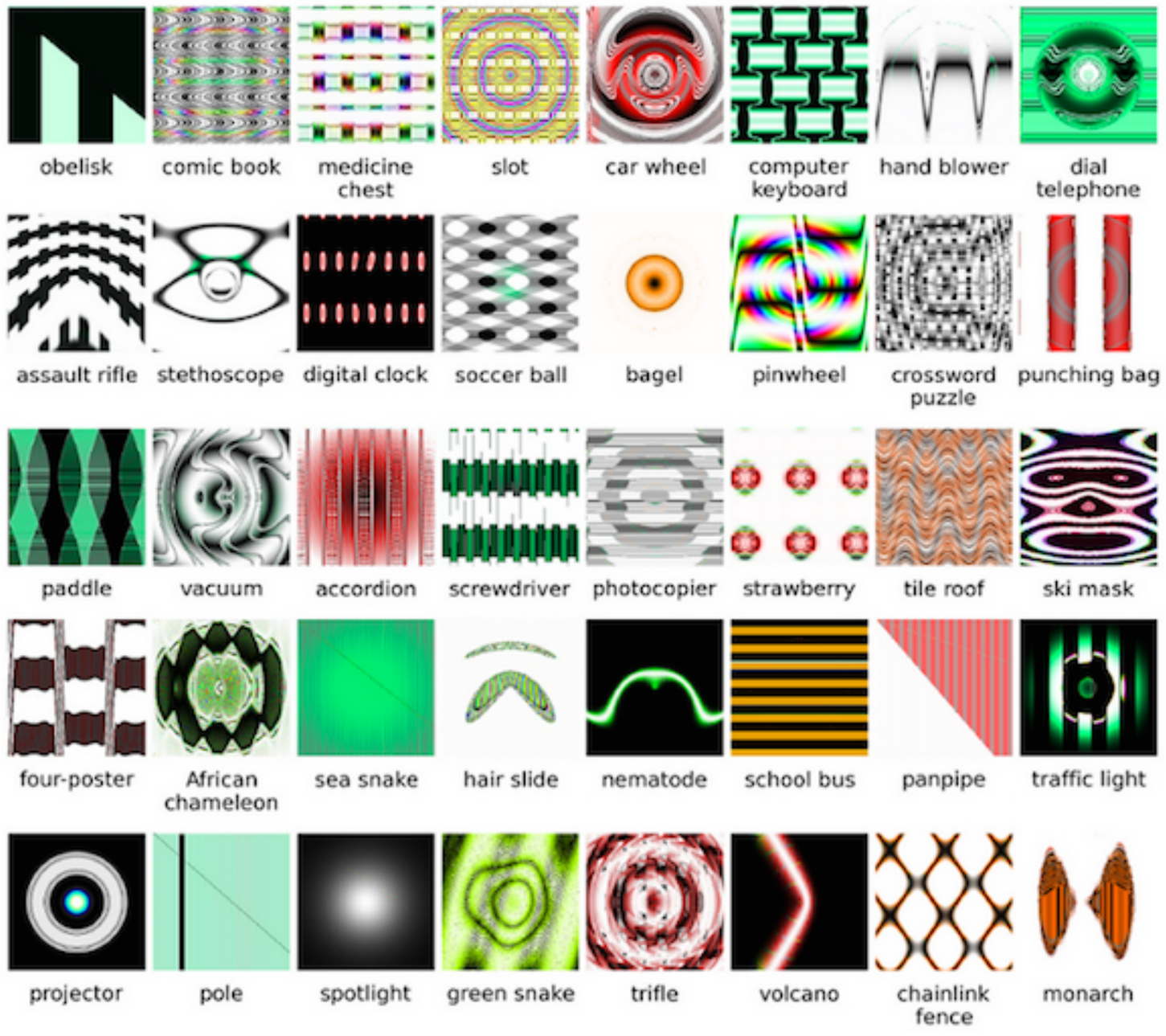
jack-clark.net

whole.

But aren't CNNs actually quite similar to the brain???

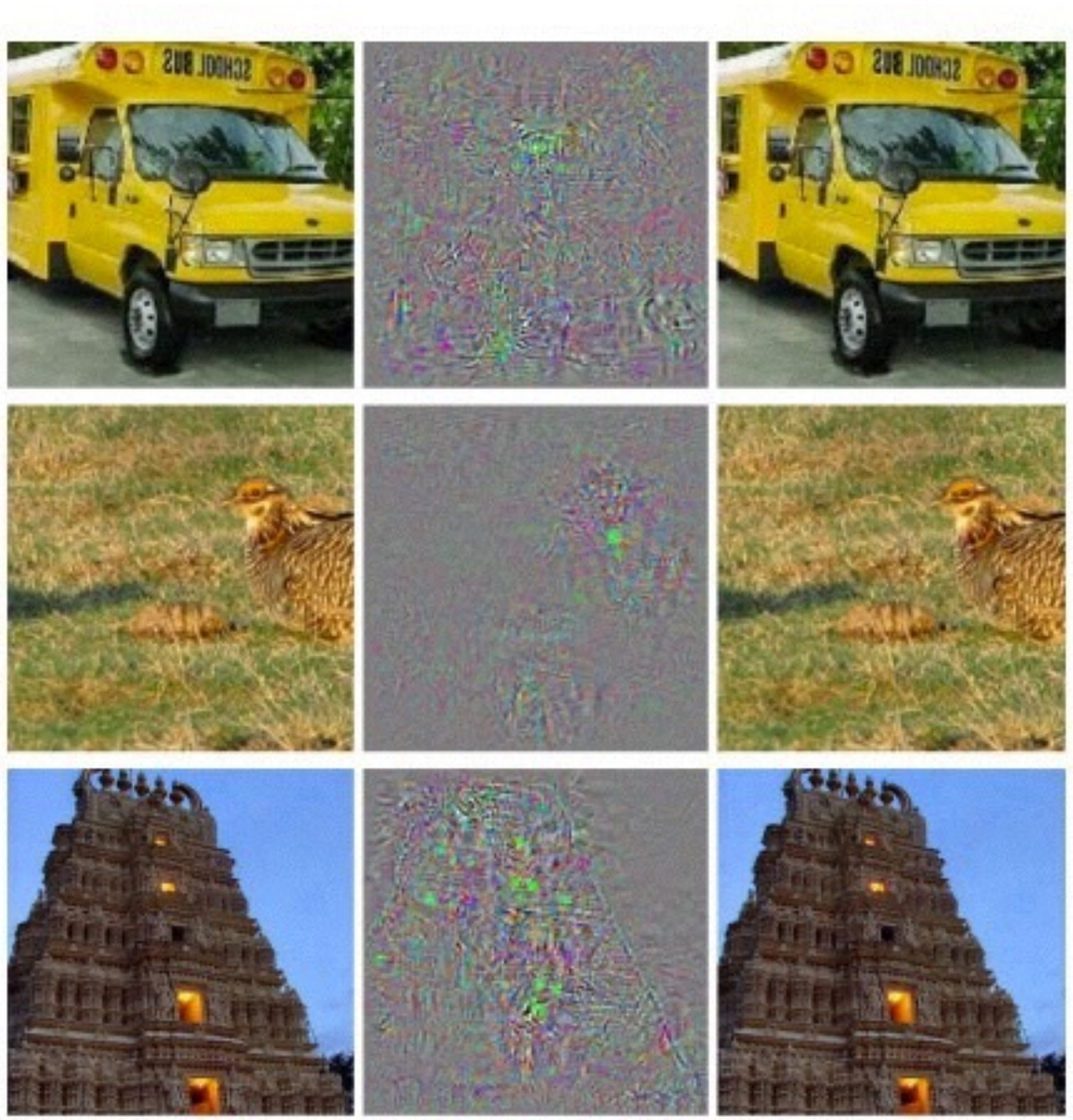
https://www.re-work.co/blog/deep-learning-peter-sadowski

Adversarial Examples



Paper: Explaining and Harnessing Adversarial Examples

Adversarial Examples



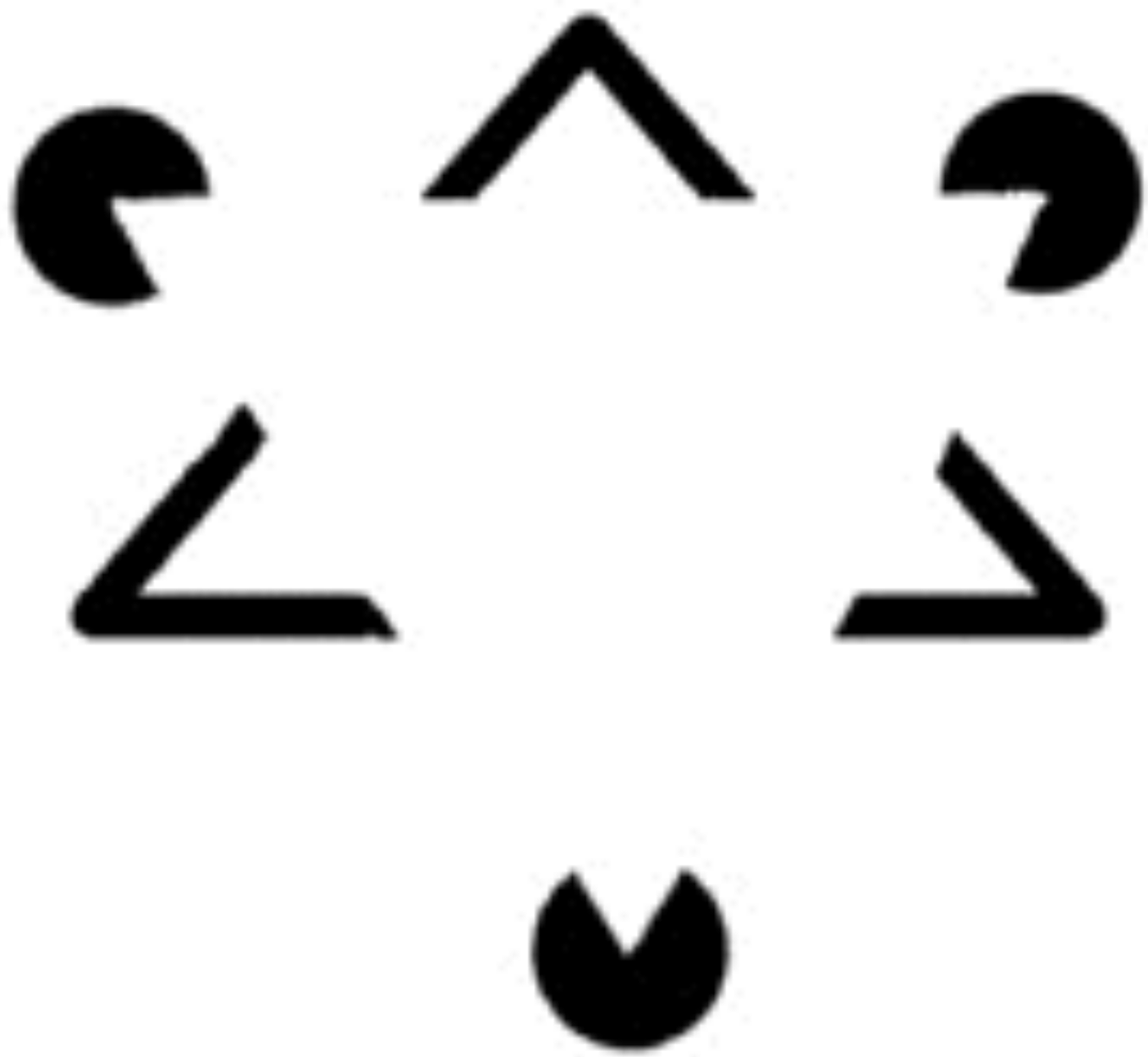
Paper: Intriguing properties of neural networks

Adversarial Examples

More training data?!

Paper: Intriguing properties of neural networks

Gestalt Law of Closure



Conclusion

Pros

- no feature engineering necessary
- setting state of the art for
  - computer vision,
  - speech recognition and
  - some text analysis
- full potential still unknown
  - they get better with more data

Cons

- doesn't work out of the box
- many choices
  - architecture
  - hyper parameter
  - learning algorithm
- needs huge amounts of data
- computational expensive

➡ lots of expensive experiments

Thank you